Tweeting across hashtags: overlapping users and the importance of language, topics, and politics

Marco Toledo Bastos
London School of Economics
University of Sao Paulo
2 Portsmouth St, London WC2A, UK
+44 2029556651
m.toledo-bastos@lse.ac.uk

Cornelius Puschmann
Humboldt University of Berlin
Unter den Linden 6
Berlin D-10099, Germany
+49 3020934309
puschmann@ibi.hu-berlin.de

Rodrigo Travitzki
University of Sao Paulo
Av da Universidade 308, Sala 18
São Paulo 05508040, Brazil
+55 1130913519
travitzki@usp.br

ABSTRACT
In this paper we investigate the activity of 1 million users tweeting under 455 different hashtags related to a wide range of topics (political activism, health, technology, sports, Twitter-idioms). We find that 70% of users in the sample tweet across multiple information streams, frequently engaging in what could be described as serial activism. We furthermore determined the dominant language in each hashtag to trace which users overlap between the thematic and linguistic communities delineated by different information streams. Although social media is frequently assumed to bring together people of different nationalities and cultures to discuss a wide range of controversial issues, our results indicate that the underlying social network that connects hashtags through overlapping users is heavily limited to linguistic and content-oriented communities. Information streams are clustered around linguistic communities, and hashtags within the same language group are clustered around well-defined topics, such as health, entertainment and politics. The only information streams that transcend language barriers are activism-related hashtags, which cluster information streams in different languages. Contrasting with the assumption that social media acts as the enabler of a globalized public debate, our results indicate a linear relationship between users who are very active in political hashtags and users who tweet across multiple political hashtags. The results suggest that activist campaigns based on social media are driven by a relatively small number of highly-active, politically engaged users.

Categories and Subject Descriptors
J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms
Human Factors

Keywords
Twitter, Hashtags, Public Space, Users, Language, Activism

1. INTRODUCTION
Social media are frequently framed as enablers of public debate, bringing people together to discuss a wide range of controversial issues. Activism in platforms such as Facebook and Twitter has been studied in the context of major events that have been widely covered by the global news media, such as the Spanish 15M (*Indignados*) protests, the Occupy movements, the so-called Arab Spring, and the European financial crisis. While social media platforms appear to be valuable tools for activists and a number of different functions they can play for online activism have been discerned, research on the complex relationship of different user communities and topics is still forthcoming. Furthermore, there is evidence that optimistic assumptions about the diversity of actors and discourses in social media are not always well-founded [3]. Opinion leaders emerge in online communities and establish themselves by being highly active and by occupying a privileged position in the social network.

In this paper we examine the role of language and topic-structure in Twitter conversations by looking into the number of users that overlap between Twitter hashtags. Our results indicate that the communities of Twitter users are structured around specific languages and topics. The only information streams that overcome language barriers are activism-related hashtags, which clustered information streams of different languages. The results reported in this paper suggest a linear relationship between users that are very active in political hashtags and users that tweet across multiple political hashtags, thus supporting the hypothesis that activist campaigns based on social media are driven by a relatively small number of highly-active, politically engaged users.

2. DATASET
The dataset comprises the activity of 3,758,160 non-unique users, of which 2,641,592 overlap between hashtags and 1,116,568 are users that participated in one information stream only. The number of users that overlap exceeds the number of unique users in a proportion of 70% to 30%. The dataset comprehends 8,449,382 (8.4 million) tweets, of which 2,848,717 (2.8 million) are retweets and 535,706 are mention-messages. There are 2,330,024,144 (2 billion) followees and 25,836,560,157 (25 billion) followers, which is consistent with previous investigations that found Twitter's distribution of followers (social graph) to be highly skewed with a low rate of reciprocated ties [4, 7].

Figure 1 shows the language division in the dataset. From a total of 455 hashtags, 244 have English as the prevailing language (53%) and 204 hashtags have non-English languages as the prevailing idiom (47%). From the 204 non-English hashtags in the dataset, 123 have Portuguese and 61 have Spanish as their predominant languages. Arabic, Dutch and Italian account for two hashtags each, while Swedish, Indonesian and Japanese are each the main language in one hashtag. Despite the limited literature and the scarcity of methods to analyze language use in social

24th ACM Conference on Hypertext and Social Media
1–3 May 2013, Paris, France.

media, previous research [2] have investigated 62 million tweets and found that English tweets account for 51% of Twitter data stream, while Japanese, Portuguese, Indonesian, and Spanish account together for 39%. The proportion of English and non-English tweets reported in previous studies are similar to those described in this study.

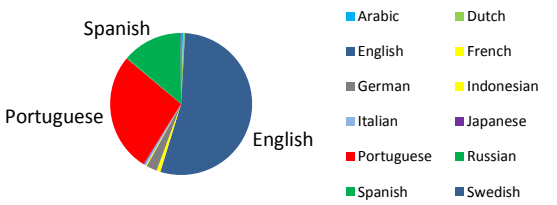

Figure 1: Languages diversity in the dataset

We mined user's activity in the 455 information streams (hashtags and keywords) and computed user's number of messages, number of followees and followers, retweets and mention-messages. We compared the activity of 3.7 million users across the 455 information streams by separating active and passive users, passive users being those mentioned by other users in conjunction with a hashtag, possibly without their knowledge. We calculated the number of overlapping users taking into consideration only users that actively posted tweets with a hashtag, excluding passive users. Table 1 shows the differences between the two sets of users in the 455 hashtags.

Table 1 indicates that highly-followed users participated mostly passively across information streams. The dataset with active users contains only a fraction of the number of followers included in the global information stream (from 25 billion to 4.5 billion followers), given that most highly-followed Twitter accounts are not present in the dataset of active users. The Twitter accounts of Barack Obama (@barackobama), media pundit Keith Olbermann (@keitholbermann), and teen-pop celebrity Justin Bieber (@justinbieber) appear prominently in the tweets of others users, who strategically associate them with causes they are most likely unaware of. We refer to this phenomenon as *sockpuppeteering*.

	All Users	Active Users	Percentag
Hashtags	455	455	100%
Users	3,758,160	3,486,317	93%
Overlapping	2,641,592	2,466,546	93%
Unique Users	1,116,568	1,019,771	91%
Followers	25,836,560,157	4,533,722,838	**18%**
Followees	2,330,024,144	1,853,827,780	80%
Retweets	2,848,717	2,584,971	91%
Mentions	535,706	535,706	100%
Tweets	8,449,382	8,449,382	100%
User Tweets	28,574,106,349	25,678,524,489	90%

Table 1: Differences between datasets with active and passive (sockpuppets) users in the 455 information streams

2.1 Sockpuppets

Table 2 shows the top 10 users by number of overlapping hashtags in the global dataset, indicating that highly-followed users (highlighted) participated only passively across the information streams. After filtering out passive users, we found that the number of retweets and mention-messages was not significantly altered, and that the number of overlapping users across hashtags was actually higher after non-active users were excluded from the dataset (71% as opposed to 70%).

User	Number of mentions in hashtags
favstar_pop	106
pierrepetrelli	82
barackobama	81
personalescrito	76
darealmaozedong	72
soniabouzas	72
waldeterossi	70
occupywallstnyc	66
eigensinn83	65
mmflint	65

Table 2: Top 10 active and non-active (highlighted) users by number of overlapping hashtags

We found that a number of media outlets, political institutions and organizations are object of sockpuppeteering. Twitter accounts of Spanish movements Democracia Real Ya! (*Real Democracy Now!* - @democraciareal) and Acampada Sol (*Occupation at Sol Square* - @acampadasol), American political movements Occupy US (@occupy_usa) and Occupy Wall Street (@occupywallst), and international political organizations Take The Square (@takethesquare), Anonymous (@youranonnews), and Tibet Truth (@tibettruth) participated only as sockpuppets in the hashtags. Twitter accounts or media outlets CNN (@cnn), Reuters (@reuters), New York Times (@nytimes) and Huffington Post (@huffingtonpost) were also highly mentioned and retweeted without having actively participated in any information stream.

This indicates that highly-followed, established Twitter accounts are mentioned across a variety of information streams as a means to attract media coverage to political events and draw attention to a particular cause or opinion. Because of the impact of highly-followed Twitter accounts, we evaluated the relative importance of such user-hubs by comparing the hashtag degree with and without sockpuppet Twitter accounts. We found no significant difference between the topology of the networks, and hashtags that host large numbers of overlapping users are not affected by the removal of non-active accounts.

3. METHODS

We collapsed the 455 hashtags into one dataset of 3.7 million active users and identified which users participated in which hashtags. We found that 70% of users tweeted across multiple hashtags, while 30% of users tweeted in only one information stream. We calculated the number of overlapping users per hashtag ($\bar{x}$=7,662 $\tilde{x}$=3,015) and the number of overlapping hashtags per user ($\bar{x}$=1.3 $\tilde{x}$=1). Table 3 shows the differences between the top 7 hashtags by network degree and by number of overlapping users in the network. Only the hashtag *occupyoakland* appears on both lists, suggesting it is both strongly connected to other hashtags (particularly the Occupy-family) and widely used by serial activists engaging in political causes.

Hashtag	Degree	Hashtag	Overla
occupyoaklan	228905	breakingdown	59843
occupytogethe	203177	grandtheftautomemories	59687
occupyla	192594	sheenroast	58754
occupydc	191005	kony2012	53852
occupyboston	187165	no1likesubecause	53372
occupydenver	173343	relationshipsendbecause	51963
occupysf	173077	**occupyoakland**	50109

Table 3: Top hashtags by network degree and by number of overlapping users

We used the co-occurrence of users across hashtags to create an affiliation (incidence) matrix that was subsequently transformed into a bipartite graph of hashtags with common users. Although hashtags with many tweets, such as *occupyoakland*, present a higher-than-average number of overlapping users in comparison to smaller hashtags, we found the topic of the conversation to be a better predictor of the number of overlapping users. This applies particularly to humorous Twitter-idiom hashtags such as *relationshipsendbecause* and *grandtheftautomemories* Figure 2 depicts the number of tweets in each hashtag by the circle circumference, showing the distribution of hashtags by total number of users and the number of overlapping users. Hashtags are equally distributed in regard to number of tweets and number of users, indicating that the number of messages or users is not of decisive importance to the number of overlapping users.

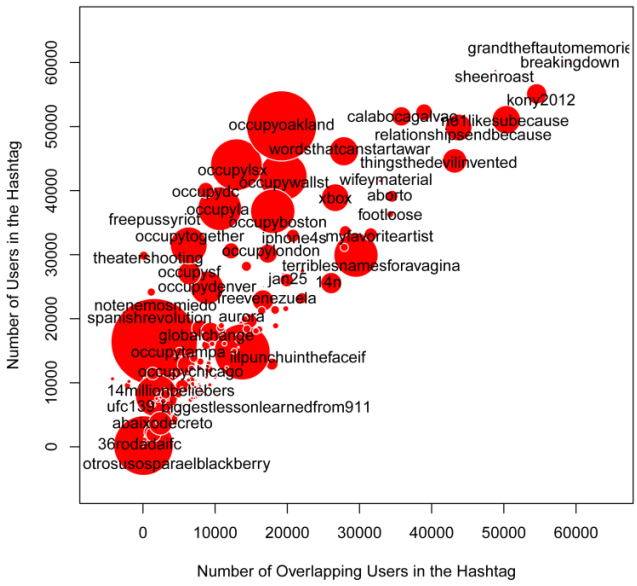

Figure 2: Distribution of hashtags by number of tweets, number of users and number of overlapping users

We calculated the number of messages tweeted per user in each hashtag to create a table of user messages versus information streams. We processed this information to compute the co-occurrence of hashtags per user in an incidence matrix of users and multiple hashtags. Lastly we created an adjacency matrix based on hashtag co-occurrence. Overlapping users were assigned as edges between hashtags, resulting in an adjacency matrix of 455 hashtag based on overlapping users between hashtags. The adjacency matrix rendered an undirected network of 455 hashtags identified according to the prevailing language used on tweets. This classification can be considered debatable, as it is based on the language of the majority of posts indicated on the language parameter of Twitter streaming API. The analysis reported in the next section result from this network of hashtags connected by overlapping users and the prevailing language in the dataset.

4. RESULTS

Two hashtags share an edge if one user tweeted across both hashtags. Nodes are sized according to the number of tweets in the hashtag and the network is colored according to linguistic communities. We found that the resulting network was dense and presented an extraordinarily high average degree (N = 455, De = 0.54, $\langle k \rangle$ = 243). We analyzed the structure of the network by CPM (Clique Percolation Method) [5] and by subsequently applying network clustering algorithms [1, 6]. The algorithms confirmed the existence of three main clusters containing the vast majority of the hashtags that match the prevailing linguistic communities within the network. The network graph indicates that the underlying social network connecting hashtags by overlapping users is heavily restricted to linguistic and content-oriented subcommunities. Information streams are clustered around linguistic subcommunities, and information streams within the same language group are clustered again around well-defined topics, such as health, entertainment and politics. The only hashtags that transcend language barriers were found to be activism-related information streams, which clustered various information streams in a number of different languages.

4.1 Languages

Co-occurrence of users across hashtags reveals how linguistic barriers impose significant divisions between Twitter users. Figure 3 shows a visualization of the network and three distinct communities of hashtags that correspond to the prevailing languages used in the dataset. Red nodes comprehend information streams mainly posted in Portuguese, blue nodes include English-hashtags only, and green nodes refer to Spanish-related hashtags. The linguistic division plays an important role in structuring the network communities. Figure 3 shows the importance of linguistic communities to the network topology, as there is little intersection between different linguistic groups. Misplaced nodes represent a small group of hashtags in which users tweeted abundantly in more than one language. These hashtags appear in the graph in a color different than the hashtags around them, as shown by the hashtag antesdelfindelmundo (*before the end of the world*).

Misplaced hashtags also indicate that hashtags can group multiple linguistic subcommunities. We assigned a unique language ID to each information stream, so hashtags in which users tweeted using different languages are prone to mismatch. Even though the hashtag aborto (*abortion*) was originally identified as part of the subcommunity of Spanish speakers, the network clustering algorithms reported that most of the overlapping users belonged to the Portuguese subcommunity. The same deviation can be seen in the hashtag antesdelfindelmundo, and particularly in meme, which has distinct meanings in English and French. Table 4 shows the clustering of hashtags according to language and topics, and groups are formed as an outcome of language subcommunities (Clusters 2 and 3) and conversational topics (Cluster 1).

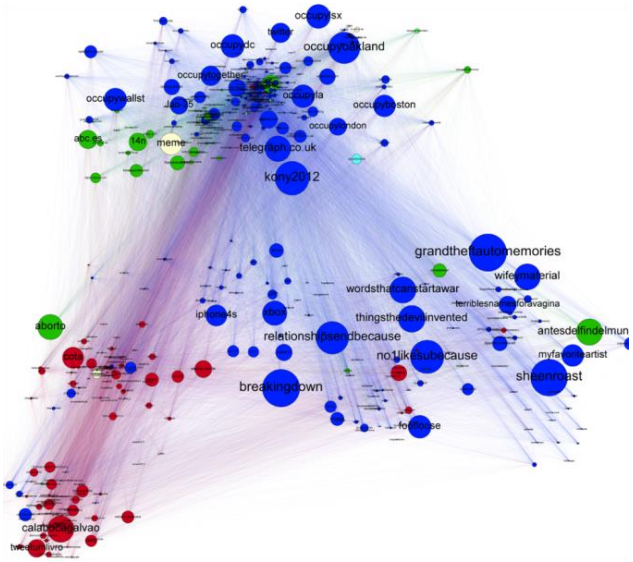

Figure 3: Network of hashtags and linguistic subcommunities

Cluster 1		Cluster 2		Cluster 3	
English	51%	English	96%	English	17%
Portuguese	1%	Portuguese	4%	Portuguese	78%
Spanish	48%	Spanish	0%	Spanish	5%

Table 4: Network subcommunities and network clusters

Common to the misclassified hashtags is the large number of tweets in different languages. The hashtags UFC126, UFC129, and UFC132 did not present a conclusively prevailing language across the information streams. These hashtags refer to "Ultimate Fight Championship," a martial art tournament in the U.S. largely ignored in countries other than the United States and Brazil. Twitter reports that most tweets were posted in English, though the total number reported is very close to the number of tweets in Portuguese. While the nodes were initially assigned as part of the English subcommunity group, the network clustering algorithm positioned the hashtags around Portuguese information streams.

4.2 Topics

As shown in Figure 3, the upper cluster in the network includes hashtags related to the political events in Spain known as 15M (*Indignados*) and the Occupy demonstrations across the United States, so that English and Spanish topics bring together different linguistic subcommunities. The underlying social network connecting hashtags by overlapping users is not only linguistic-oriented, but also content-oriented, particularly in regard to politics and activism-related events such as the 15M and the Occupy demonstrations.

In order to evaluate network communities defined by different topics, we first focused on the English information streams. In Figure 4 we removed all non-English nodes and pushed important topic subcommunities to the periphery of the network. Nodes are sized according to the number of tweets in the hashtag and colors identify different topics. As shown in Figures 4 to 6, English hashtags cluster around clearly-defined topics. Health-related hashtags cluster with other health-related hashtags at the bottom of the graph. Political hashtags cluster with other politics-related hashtags in the left corner of the graph, particularly Occupy-related hashtags. There are five large subcommunities of topics: entertainment, politics, technology, Twitter-idioms, and health.

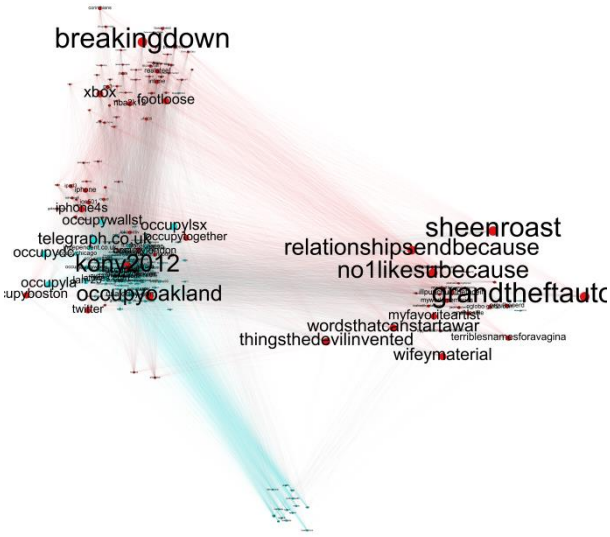

Figure 4: English hashtags and topic subcommunities

Figure 4 shows sport-related hashtags below the politics group, and above we find technology-related hashtags (Figure 6). The cluster on the top of Figure 4 revolves around entertainment (Figure 6), with game-related hashtags on the left and movies and television on the right. The remaining large cluster on the right side of the graph includes a large number of Twitter-Idiom hashtags, in which users concatenate common words into neologisms that serve as a marker for playful and ironic conversational themes.

We analyzed Portuguese and Spanish information streams and found a similar division between topics. Figure 7 shows that Portuguese hashtags are also clustered along topics, with political and news-related information streams grouped at the bottom of the graph, and sports, music, movies, and Twitter-idioms grouped on the top. Figure 7 also shows Spanish hashtags grouped around one large cluster. The content of these hashtags mostly relate to politics, so hashtags are largely clustered around a group of tightly-knit though not particularly large hashtags used during the 15M demonstrations in Spain.

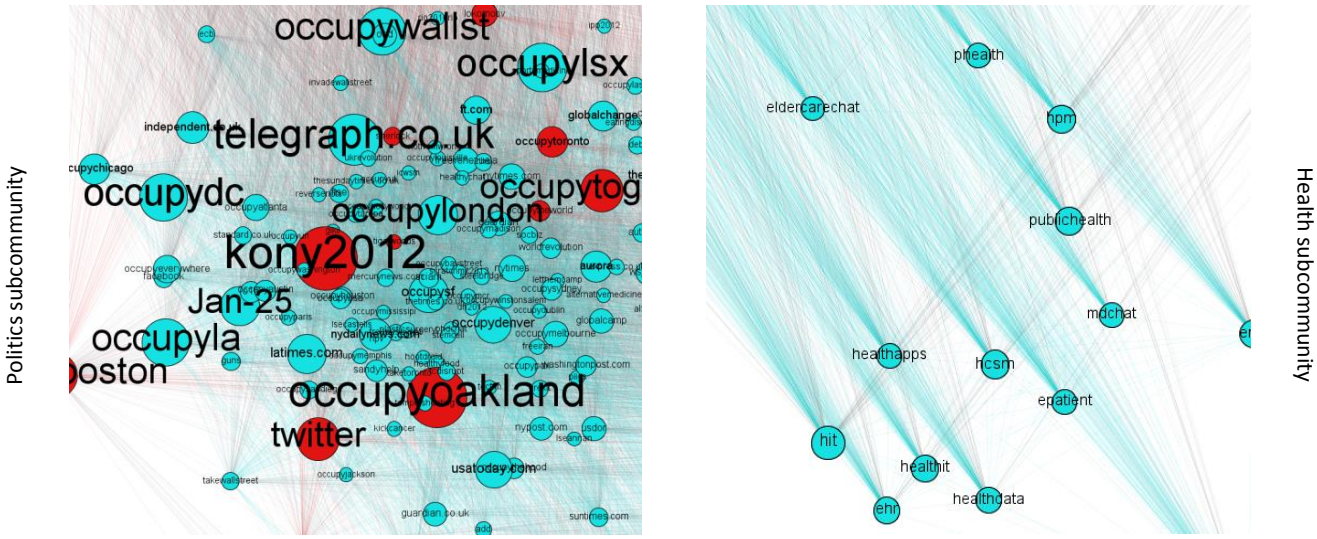

Figure 5: Detail of topic subcommunities in English hashtags shown in Figure 4

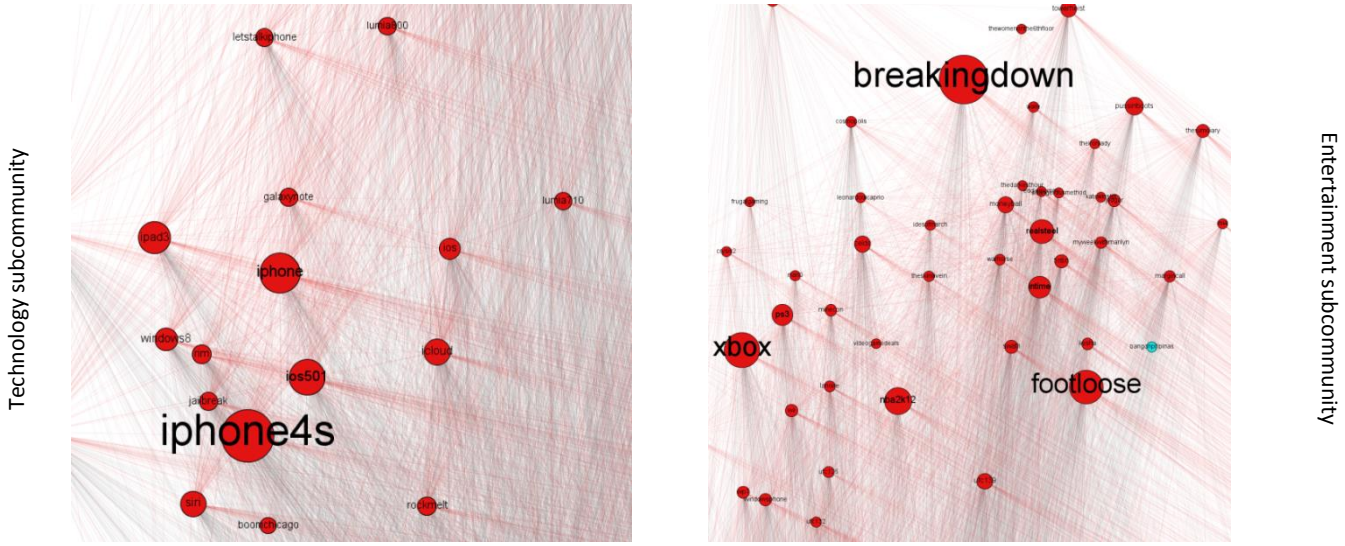

Figure 6: Detail of topic subcommunities in English hashtags shown in Figure 4

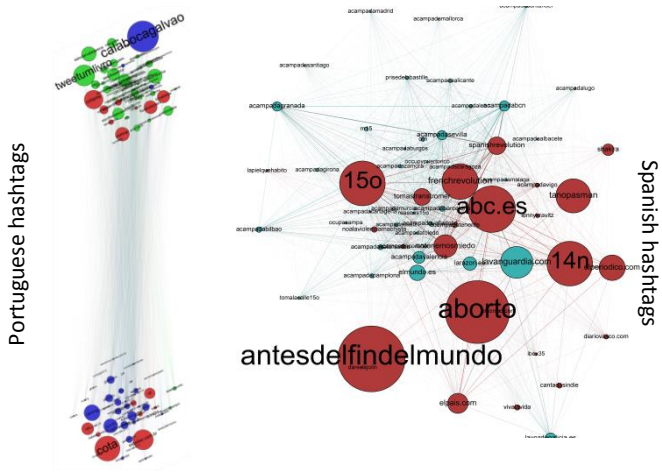

Figure 7: Topic subcommunities in Portuguese and Spanish hashtags

4.3 Politics

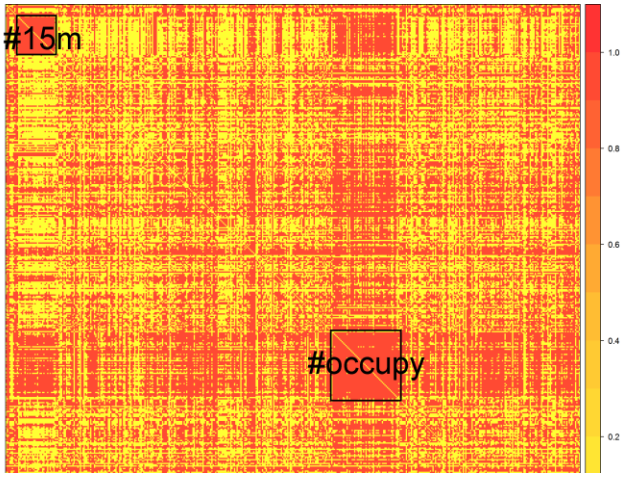

Figure 8: Overlapping of user between hashtags

The network graph shown in Figure 3 and the information presented in Table 4 indicate a relationship between users that are very active in political hashtags and users that tweet across multiple political hashtags. The results of the network clustering algorithms described in Table 4 show that Cluster 1 is divided by English and Spanish users, but operate as one marketplace of political ideas. Figure 7 shows that activism-related hashtags, particularly events connected to 15M and Occupy protests, have the highest network degree, thus suggesting that political campaigns based on social media are driven by highly-active, politically engaged users that tweet across different hashtags and are immune to language barriers. Thus, political hashtags transcend linguistic communities, grouping together users and messages produced in a number of different languages.

5. REFERENCES

[1] Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008 (10).

[2] Hong, L., Convertino, G. and Chi, E.H. Language Matters In Twitter: A Large Scale Study, 2011.

[3] Jungherr, A., Jürgens, P. and Schoen, H. Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment". Social Science Computer Review, 30. 229-234.

[4] Kwak, H., Lee, C., Park, H. and Moon, S., What is Twitter, a social network or a news media? in WWW 2010: Proceedings of the 19th international conference on World Wide Web, (New York, NY, USA, 2010), ACM, 591-600.

[5] Palla, G., Derenyi, I., Farkas, I. and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435 (7043). 814-818.

[6] Rosvall, M. and Bergstrom, C.T. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 105 (4). 1118-1123.

[7] Wu, S., Hofman, J.M., Mason, W.A. and Watts, D.J. Who Says What to Whom on Twitter, Yahoo! Research, New York, 2011.