Birds of a Feather Tweet Together: Computational Techniques to Understand User Communities in Social Networks

David Burth Kurka
(1) University of Campinas
(2) Imperial College London
d.kurka@ic.ac.uk

Alan Godoy
(1) University of Campinas
(2) CPqD Foundation
godoy@dca.fee.unicamp.br

Fernando J. Von Zuben
University of Campinas
vonzuben@dca.fee.unicamp.br

ABSTRACT

The study of social systems shows that there is a relationship of mutual influence between social connections and individual behavior, known as homophily. In this work, we developed a methodology to allow the analysis of interests of groups of users in Twitter network, based on automatic community detection and tweets ranking. The techniques presented reveal evidences that the presence of communities is related to topic specialization, and allow the characterization of elaborate profiles of groups of users based only on their location on the network.

CCS Concepts

•Networks → Online social networks; •Human-centered computing → Social network analysis; Collaborative and social computing;

Keywords

Online social networks, social network analysis, community detection, homophily, complex systems.

1. INTRODUCTION

The experience of the last decade has shown that online social networks (OSNs) are not only useful for the amusement of Internet users, but can also be a valuable source of data for the study of social systems. The millions to billions of users who access OSNs services everyday are providing researchers an unprecedented possibility to gather information from human activity and social behavior, enabling the investigation of complex issues [7].

One of the interesting topics that can be explored in this context is the creation and diffusion of content. OSNs users are constantly involved in complex dynamics of message sharing, which may result in the emergence of new trends, collective mobilization and opinion formation. Understanding the essential mechanisms of this process can be useful to areas such as politics [6] and marketing campaigns [10, 20].

This paper explores the interplay between social connections and content shared by individuals in an online social network, which is one aspect of social communication. We investigated here how it is possible to understand behavior and interests of a group of individuals based on their social connections, paying special attention to the role of homophily – the tendency for someone to establish connections to similar peers. Analytical tools were developed – and applied to real data extracted from OSNs – to pinpoint content relevant to the understanding of interests of communities and users composing them. The techniques used are based solely on the knowledge of the users that shared each message, not resorting to their contents. We believe that, in addition to indicating the relevance of sharing information for the classification of messages, such tools can be useful to the practical study of social systems and to the development of new applications, such as recommendation systems.

The paper is structured as follows. In Section 2 we introduce relevant literature on OSN and homophily. In Section 3, we present the methods used to acquire and analyze data. Then, in Section 4, we report and analyze the results obtained using our methodology. Finally, in Section 5, we present final remarks, discussing implications of this work and possible future directions.

2. BACKGROUND

An increasing number of studies have been conducted over the last years focusing on online social networks [7]. Dynamics of information diffusion [3], public opinion prediction [19], collective sentiment analysis [9, 18] and the formation of social structures [8] are among the subjects explored.

An important aspect investigated is the presence of self-organized processes. Despite not having central controllers that rule on how content is disseminated or connections between users are created, OSNs display many organized behaviors. Common examples of such process are the collective curation of contents [17] and information diffusion cascades [5].

2.1 Communities and Homophily

Self-organizing processes also take part on how connections are formed in an OSN. Usually, social connections are not created uniformly between all individuals, but are concentrated in few *hubs* [1]. Thus, when analysing topologies of real networks, *communities* of individuals which are more likely to be linked between each other, with fewer connec-

Published as part of the #Microposts2016 Workshop proceedings, available online as CEUR Vol-1691 (http://ceur-ws.org/Vol-1691)

#Microposts2016, Apr 11th, 2016, Montréal, Canada.

tions between individuals from different communities, are found. Social systems exhibit communities in many different levels: people inside families, sharing specific common interests, or living at the same city or nation have many more connections to other in-groups than to out-groups [12].

Considering the flow of information that travels over the network, groups with highly connected individuals may imply a redundancy of communication channels, possibly enhancing information that pass through or are generated by these groups. In the case of online social networks, it can imply that a content may be more easily proliferated and reinforced inside a community after it is shared by a member of such group. While communities are expected to influence on how their members receive and process information, it is also believed that community formation is influenced by pre-existing affinities between members [21].

Researchers identified in social networks outside the virtual world a tendency (not only within communities) of individuals with common interests to be usually connected to each other [11]. Such phenomenon is called *homophily* and is observed, also, in OSNs.

Kwak et al. [8], in an early study analyzing Twitter's data, showed evidences of homophily among users with same localization and same number of friends (popularity). Also on Twitter, Wu et al. [22], found a strong tendency where users belonging to a same category (e.g artists, organizations, bloggers) would communication among themselves. Romero et al. [14] studied the relationship between the (explicit) network of friendship and the (implicit) network of topical affiliations – i.e., the communities formed by users interested in a common topic. They showed that both networks have considerable intersection and that users tend to connect to other users with common interests. This correlation allows the prediction of friendship connections from *hashtag* diffusions and also the forecast of the future popularity of a *hashtag* from the friends network.

Bollen et al. [2] verified that users' emotions is strongly correlated with social connections, showing that users considered happy tend to be linked to each other. Salathé et al. [16] explored how a network with signs of homophily interfere on the spread of sentiment towards a new vaccine, showing how negative opinions can be reinforced in such environments.

2.2 Communities Detection

Some of the most common approach to community detection are modularity-based algorithms [13], which look to partition a network in communities so that the summed weight of all connections between two communities is minimized. This approach, as most other techniques usually applied for community detection, however, is insensitive to direction of connections in the network. In systems where patterns of flow among individuals are relevant, however, ignoring connection direction may disregard information valuable to the comprehension of collective behavior.

In order to address this issue, Rosvall and Bergstrom [15] proposed a flow-based method, which defines a community as a set of individuals "among which information flows quickly and easily". Their algorithm takes advantage of both direction and weights of connections, using information theoretical measures and a random walk as a proxy of information flow. Using a greedy search, they look for a partition of individuals that define a two-level description – the first level indicating the partition and the second indicating a individual inside the partition – that minimizes the expected description length of a random walk in the graph. This partitions are, thus, the communities of individuals in the network.

3. METHODS

3.1 Data Acquisition

As Twitter has plenty of public data available online, it is a good source for creating a database of social events. However, the rate limits imposed by its API[1] hamper the download of large volumes of data from events that took place in the past. An alternative is to use Twitter's Streaming APIs, which allows the download of messages in real time, as they are posted.

Therefore, in order to collect a satisfactory amount of data to be analyzed, we decided to track the interactions between a popular user account producer of original content and the users that share (i.e. retweet) these messages. As popular Twitter accounts interact with many users daily, this approach revealed to be an effective way for collecting message diffusion processes and user interactions as they happen.

We chose the Brazilian largest newspaper Twitter account, Folha de São Paulo[2], and collected: original messages posted, retweets of those messages posted by other users, account details of those users and the relationships (followers and followees) of all of them.

3.2 Automatic Topic Classification

As with many other newspapers, almost all messages published by the chosen source are headlines, followed by a link to the newspaper's website with the news' full content. As the news articles on the website belong to thematic categories (newspaper's sections), it was possible to automatically attribute a class to each tweet, based on these categories. This procedure was carried out to all tweets and six most common topics were verified, namely: "daily life news", "sports", "world", "politics", "entertainment" and "market".

3.3 Detecting common interests in communities

After collecting and classifying all data, we then evaluated whether retweeting behaviors of communities' members are coherent among themselves regarding the subjects of the shared messages. In order to detect groups of tightly connected users from the social connections observed, we executed Rosvall and Bergstrom's community detection algorithm [15] in our social network, an algorithm focused on locating groups of users among which information can flow more quickly.

Understanding the interests of a community is a hard task. To address such issue, we propose here an adaptation of a well-known statistics from text mining, the term frequency-inverse document frequency (*tf-idf*) [4].

The tf-idf method usually considers a corpus of documents, each composed of different terms. By comparing the frequency of specific terms inside and outside documents, values are assigned to each term, pondering its importance

[1] https://dev.twitter.com/rest/public/rate-limiting
[2] https://twitter.com/folha

for each document. The tf-idf of a term t in a specific document d is calculated as follows:

$$tf\_idf_d(t) = tf_d(t) \times idf(t),$$

where $tf_d(t)$ is a value that is higher the higher the frequency of the term t in d while $idf(t)$ is inversely proportional to the frequency of t in all documents of a corpus.

In the traditional use of the tf-idf algorithm in text-mining, a term is a word and a set of terms is a textual document (e.g. a book, or a webpage). This creates a *document-term matrix*, where each row represents a document, each column a word and each cell, indexed by (i, j), the frequency of the term j in document i. For the case of our application, we adapted this approach by considering each tweet as a term and each community as a document. Therefore, we built a *community-tweet matrix* where each cell (i, j) represents the number of retweets of tweet j in community i.

Thus, when the same operations of tf-idf are applied to a community-tweet matrix, tweets that are both highly shared by members of a community and more akin to such community's behavior are highlighted with higher scores than the others.

4. RESULTS

4.1 Database Description

From March 19, 2014 to September 21, 2014, all messages (tweets) posted by the source account, as well as any share of this content by other users (retweets) were collected.

It was possible to track, from the data collected, a large amount of information diffusion processes triggered by the observed source. During the observed period, 13463 distinct and original messages posted by the source account were collected. From this set, a series of filters were sequentially applied, forming a more appropriate dataset for the work, as described below:

- Filter 1 - Only messages which had received at least 20 retweets were selected, resulting in a group of 4671 distinct messages;
- Filter 2 - Messages not belonging to one of the six main categories ("everyday news", "sports", "world", "politics", "entertainment" and "market") were removed from the above set, resulting in a group of 3185 messages;
- Filter 3 - Users who retweeted more than 60% of the publications were considered automated scripts (bots) and therefore removed from the database and from the retweets count (just one user was identified as such);
- Filter 4 - As in 2014 Brazil held the Football World Cup and presidential elections, there were a high number of messages in categories "politics" and "sports". In order to balance the proportion of messages in each topic, a maximum limit of 450 messages was set for each category.

The filtered data resulted in a collection of 2444 messages, of which 44320 distinct users retweeted one or more messages at some moment. From the users lists of followers and followees, it was possible to characterize a network, registering the social connections between them. The source user, Folha de São Paulo, was removed from the network, so that the interactions among its followers became the focus of the analysis. Table 1 presents a more complete characterization of the collected data and the network formed between users. It is worth pointing, also, that most users do not participate actively on the diffusions, implying in high diversity of users participating in the processes, but low recurrence: during the observation period, each user retweeted in average only two messages from the source.

Table 1: Characterization of the collected data.

Number of users	44320
Connections between users	673982
Average degree	30.42 $(\langle k_{in} \rangle = \langle k_{out} \rangle = 15.21)$
Most followed user	@rodrigovesgo (Comedian) – 5742 followers
Most popular message	743 retweets
Clustering coefficient	3.66% (C for a randomized network is 0.07%)
Diameter	15
Average path length	4.29
Strongly connected components	12358
Size of giant component	31493 (71% of all users)
Total number of retweets	110389 (2.49 per user / 45.16 per message)

4.2 Collective Coherence in Communities

After running the community detection algorithm, 4278 communities were detected, many with 2 elements (2 connected users, isolated from the rest of the network), but 26 larger groups with 200 users or more have also been identified.

A first experiment involving communities consisted of checking how coherent communities behaviors were, by comparing the frequency at which specific messages were shared inside communities and in the complete network. For each community with a relevant number of users (200 or more), it was computed the number of retweets for each Folha de São Paulo's original tweet. A high coherence was verified in the behaviors of individuals belonging to the same community. Table 2 compares the sharing rates inside and outside the four largest communities, showing the five most distinctive cases where the community has singular sharing behavior, differentiating from the global behavior.

The cases shown on Table 2 indicate that there is a certain level of coordination in the selection of shared messages by each community. It is interesting to see how some messages are much more emphasized inside a community, than it was in the general network. Community 2, for example, present higher sharing rates for its messages, compared to the rest of the network. Equally interesting are cases where the community seems to suppress the spread of a message, as seen in community 8, where highly retweeted messages are less emphasized by the community's members.

For comparison purposes, an attempt to break the relationship between user connections and postings was made, by randomly swapping the retweet pattern of different users on the database, while keeping their social connections. Thus,

Table 2: Comparison of the frequencies of retweets between inside and outside the four largest communities.

Community 4, members: 4870, total tweets: 8978					
	1st	2nd	3rd	4th	5th
Inside	0.7%	1.0%	0.8%	0.1%	0.3%
Outside	1.8%	1.8%	1.5%	0.6%	0.8%

Community 1, members: 2621, total tweets: 12292					
	1st	2nd	3rd	4th	5th
Inside	2.5%	1.9%	2.1%	1.7%	1.4%
Outside	0.6%	0.3%	0.6%	0.2%	0.0%

Community 2, members: 1022, total tweets: 3160					
	1st	2nd	3rd	4th	5th
Inside	4.9%	3.7%	3.7%	3.6%	3.4%
Outside	0.2%	0.1%	0.2%	0.4%	0.4%

Community 8, members: 742, total tweets: 2117					
	1st	2nd	3rd	4th	5th
Inside	0.1%	0.0%	0.9%	0.8%	1.1%
Outside	1.7%	1.4%	0.1%	0.1%	0.3%

under this new experiment, the communities remain defined as they were originally, but their retweets have patterns from users of different communities, in order to eliminate any homophily related to retweeting behavior but keeping intact other characteristics of our data (as global retweet counts and correlations between sharings by individual users). The same analysis made before is now performed on the randomized dataset, resulting on Table 3. It becomes evident that when homophily is suppressed, communities lose their particular behavior and present sharing rates closer to the whole network rates, indicating that differences in retweeting patterns are not only artifact of communities' finite sizes.

Table 3: Comparison of the frequencies of retweets between inside and outside communities for the randomized dataset.

Community 4, members: 4870, total tweets: 12214					
	1st	2nd	3rd	4th	5th
Inside	1.3%	0.8%	0.9%	0.2%	0.6%
Outside	1.7%	0.5%	0.7%	0.5%	0.4%

Community 1, members: 2621, total tweets: 6967					
	1st	2nd	3rd	4th	5th
Inside	1.3%	0.5%	0.6%	0.4%	0.5%
Outside	1.0%	0.2%	0.3%	0.1%	0.2%

Community 2, members: 1022, total tweets: 2505					
	1st	2nd	3rd	4th	5th
Inside	1.8%	1.4%	1.1%	0.9%	2.0%
Outside	1.0%	0.6%	1.7%	0.3%	1.4%

Community 8, members: 742, total tweets: 1619					
	1st	2nd	3rd	4th	5th
Inside	0.9%	1.1%	1.1%	0.7%	1.2%
Outside	0.2%	1.7%	0.5%	0.1%	0.7%

To verify how strong was the preference of members of a community for specific messages, we compared the Gini index[3] for their number of retweets within each community in the real data with the same index for the data in the randomized dataset. Our results indicate that the real communities have stronger preferences for specific messages, with satisfactory statistical significance when considering all communities (Wilcoxon test, $p = 1.23 * 10^{-15}$, Gini index difference of $1.5 * 10^{-4}$) and even more pronounced when restricting the comparison only to larger communities, with at least 50 individuals (Wilcoxon test, $p = 2.13 * 10^{-4}$, Gini index difference of $8.26 * 10^{-3}$).

4.3 Topic Specialization

A second experiment consisted in the analysis of how topics are distributed among communities. For this, the six standard categories were considered and the number of messages of each category shared by each community was computed. First, Figure 1a shows the general distribution of messages per topic for the whole network. Then, Figures 1b-1h show the distribution for seven distinctive communities chosen among those with more than 200 members. The figures demonstrate how communities' topic distribution may have different profiles compared to the rest of the network.

From the graphs presented, communities 1 and 16 (Figures 1b and 1f) seems to have a stronger interest in political issues, with community 1 showing an interest in daily life news slightly above the average. Community 12 (Figure 1e) also shows a higher interest in politics, but divide it with a focus on sports news. Community 21 graph (Figure 1g) shows a preference for daily life news, but not in a very distinctively way. Community 10 (Figure 1d) does not have a distribution very different from the whole network (Figure 1a), being a representative of the average preferences. Community 25 (Figure 1h), in turn, is remarkably different, with over 50% of its retweets being about entertainment and practically all the rest regarding sports news, almost ignoring the other topics.

4.4 Detecting Relevant Messages

By applying the *tf-idf* normalization on the retweet counts of the communities we identified the most characteristic tweets for each community. Table 4 presents the top five messages for the largest communities in the network. The messages content were translated from Portuguese to English, with some translation notes (in brackets), where necessary.

It was possible to deepen the analysis of each community profile beyond what would be possible by simply looking to the distribution of general topics in each community. Analysing each group of messages individually, it is possible to notice specific and subjective categories. For example, although communities 1 and 16 both have an emphasis in politics, community 1 seems to be supportive to the Brazilian government – focusing on good results of politics made by the government and scandals of the opposition – while 16 appears more involved in topics related to (at the time) election's opposition candidates.

A very interesting conclusion comes from the analysis of the relevant messages from community 12, as we discover that the tweets are not connected by the newspaper sections,

[3] The Gini index is a measure of how unevenly a value is distributed among elements of a group – in our case, if the Gini index is close to 1 then most of the retweets seen in a community were associate with few messages, if its value is near 0, then the distribution of retweets is closer to uniform.

Table 4: *tf-idf* results, showing most representative tweets in selected communities.

Community 1 – members: 1022, retweets: 3160

tf-idf	Category	Tweet's content
4.09	politics	Most Minas [Brazilian state] voters are unaware of airport made by Aécio [oposition candidate for presidency]. http://t.co/9pmkp1yS1P
4.02	market	Oil production in the country grows almost 15% and hit record, says ANP. http://t.co/Z9kgIRmoIh
3.70	daily life	Book about Lula [former president] will be the last of my career, says writer Fernando Morais. http://t.co/jw84KtniZw
3.67	daily life	Brazil has reduced by 50 % the number of people suffering hunger, the UN says. http://t.co/mOVqNYUGUa
3.63	market	Govern expands My House [housing program] in 350 thousand units in the first half of 2015. http://t.co/zAoTS3vjyU

Community 9 – members: 272, retweets: 636

tf-idf	Category	Tweet's content
3.79	sports	Brazilian national football team will play in the new stadium of Palmeiras [footbal team] http://t.co/V8DoEI42bm
3.69	sports	Palmeiras was born champion with Oberdan Cattani in goal. http://t.co/5th4sBM553
3.68	sports	Maurício de Sousa [Brazilian cartoonist] makes drawing in honor of Palmeiras centenary. http://t.co/3CBPgidV9O
3.51	sports	Cristaldo scores, Palmeiras beats Criciúma [footbal team] and wins 1st with Dorival. http://t.co/MJZNLxUEA6
3.35	sports	Fans flock to the streets to wait centenary of Palmeiras. http://t.co/uwdrcb5wCF

Community 10 – members: 283, retweets: 861

tf-idf	Category	Tweet's content
4.06	market	With wicked face, Harley-Davidson Fat Boy Special is sweet to drive. http://t.co/FeyVtebUX7
3.99	world	Pope Francis says corrupts will be held accountable to God. http://t.co/dd77QEnUei
3.89	market	Federal prosecutor's office of São Paulo denounces Eike Batista [business man]. http://t.co/SVQCBYPVPd
3.81	entertain	85 years-old, the cult filmmaker Alejandro Jodorowsky conquers Twitter with philosophy and mysticism pills. http://t.co/wv9A1ggxâĂę
3.75	daily life	Alckmin [São Paulo state governor] will sanction this week bill prohibiting masks in protests. http://t.co/EkMmwM7jHk

Community 12 – members: 316, retweets: 820

tf-idf	Category	Tweet's content
5.03	politics	With police strike, army and national security force will secure Pernambuco [Brazilian state]. http://t.co/7ynDAQwS8u
4.78	sports	Court declares Sport [Pernambuco's football team] as the sole champion of 87; Flamengo [footbal team] can go to the Supreme Court. http://t.co/BeolyRwnpd
4.51	politics	PE [Pernambuco state] says that it will only negotiate with PM [police] if strike is over. http://t.co/J3qL9ZBjic
3.91	sports	Suspect of trowing toilet bowl that killed fan is arrested in Recife [Pernambuco capital] http://t.co/m93KTdGagq
3.83	sports	#FolhaintheWorldCup skewer will cost R$ 15 in the World Cup stadiums. View the price of other items: http://t.co/k4F8QPJ4RP

Community 16 – members: 215, retweets: 527

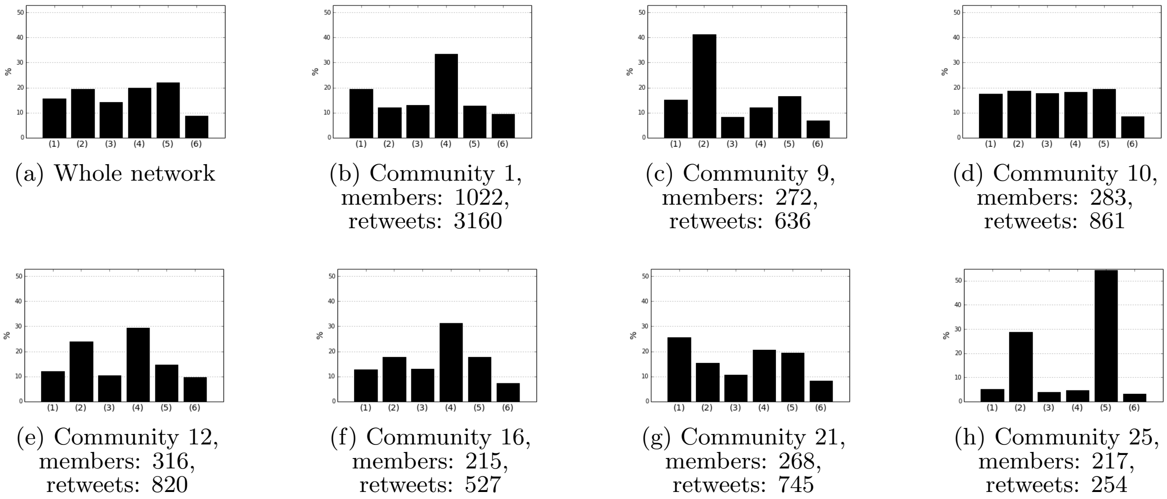
tf-idf	Category	Tweet's content
4.15	politics	Even with rain, population attends Campos' [PSB's presidential candidate dead in an airplane crash] funeral. http://t.co/1mkl0IpO4W
4.08	politics	In Maranhão [Brazilian state], Campos says that will send Sarney [politician] to the opposition. http://t.co/0qMC9uxfvw
3.85	politics	Sarney tells Dilma [Brazilian president] that he won't be a candidate anymore. http://t.co/Dks1zZmE02
3.77	politics	Frederico Vasconcelos: Leandro Paulsen: "It would be good to have a tax in the Supreme Court. http://t.co/5IPV53igmh
3.75	politics	In video recorded inside Papuda prision, José Dirceu [arrested politician] complains about the closed regime. Watch: http://t.co/7gmRr9CwLU

Community 21 – members: 268, retweets: 745

tf-idf	Category	Tweet's content
4.32	politics	SP [São Paulo state] subway workers reject agreement and decide to go on strike on Thursday. http://t.co/sxcervHS2x
3.81	daily life	Drivers and conductors block garages in Osasco and Diadema [cities near São Paulo]. http://t.co/tULggQlUDq
3.74	politics	#FolhaintheWorldCup Chamber of deputies reject holiday during Brazil's matches. http://t.co/Cuw9ue6qNa
3.69	daily life	Subway does a campaign against sexual harassment in train and subway stations. http://t.co/uVJo0mX3dD
3.58	world	Chinese ship detects signal that can be the black box of missing plane, says the agency. http://t.co/X9HP1bmUgG

Community 25 – members: 217, retweets: 254

tf-idf	Category	Tweet's content
3.81	entertain	Aged 22, a member of One Direction buys English football team. http://t.co/HoxsKSpXY9
3.06	world	Data protection in the digital age requires "increased attention", says Dilma. http://t.co/CXh5pQuCRf
3.01	entertain	American college offers lectures on Miley Cirus. http://t.co/KjtiKrKcnc(via@sitef5)
2.87	market	Banks plan to 'extinguish' DOC [money transfer format] until 2015. http://t.co/9CnHHUNES1
2.87	daily life	Demonstration blocks roads of the east side of São Paulo city. http://t.co/ix7I5kIukU



(a) Whole network

(b) Community 1, members: 1022, retweets: 3160

(c) Community 9, members: 272, retweets: 636

(d) Community 10, members: 283, retweets: 861

(e) Community 12, members: 316, retweets: 820

(f) Community 16, members: 215, retweets: 527

(g) Community 21, members: 268, retweets: 745

(h) Community 25, members: 217, retweets: 254

Figure 1: Topic distribution in the entire network and within the largest communities. In the histograms, topics are represented as follows: (1) "daily life news"; (2) "sports"; (3) "world"; (4) "politics"; (5) "entertainment"; (6) "market".

but by subjects regarding the Brazilian state of Pernambuco and its capital, Recife. All the messages presented were related to events taking place in the state, involving both politics matters (police strike) and sports (2014 FIFA World Cup events). This analysis shows both the limitations of the standard categorization of topics (the six classes defined by the newspaper) and the potential of the *tf-idf* technique on revealing new subjective connections among messages.

Another strong topic specialization is noticed in community 9, where all the top five messages are related to the football team Palmeiras, giving evidence that the community consists mainly of the team's supporters. Community 25 is specialized in entertainment topics, showing an apparent tendency to emphasize messages related to international pop culture. Interestingly, although the topic distribution in this community was predominantly on entertainment and sports, the *tf-idf* normalization reveals the relative relevance of messages in other categories, such as world and market (market is the least shared topic among the six categories). Sports tweets were not present among the top five messages, which can probably mean that the sports messages shared by the community followed the general distribution, not revealing a distinctive behavior of the community.

This kind of qualitative analysis of communities behavior could be made with most communities detected in the network, but are not presented here, due to space constraints. Other examples of topic specialization present in communities include: regional news (from diverse Brazilian states), international politics, economy, football discussions (in general and regarding specific teams) and corruption.

5. DISCUSSION

This research presents a computational framework for general investigations on collective behavior. When applied to a large dataset, the method presents new evidences of homophily in Twitter's network. Despite the existence of homophily in Twitter was already found in different studies [2, 8, 22], homophily may be based on many different criteria, as ethnic background, social class, mood, etc. The results we presented highlight the relation between shared interests and Twitter's structure.

The use of *tf-idf* jointly with community detection was able to group and order messages according to their relevance to a social community, enabling the characterization of complex behavior profiles inside communities. The presented method was able to reveal more nuanced classes of contents, such as political positions, regional matters, fan clubs, that were not covered in the original six categories, defined by the newspaper's staff with the specific purposes of organization and classification. It is relevant to notice that, beyond Twitter, the same technique can be applied to the analysis of different sets and databases from social networks, enabling similar studies for different services and contexts.

The empirical results of this study show new concrete evidences of how individual behavior and social connections are closely related, as expected in theory. In fact, so much information is present in social connections that it was even possible to find groups of similar messages without any knowledge of their content, using techniques that only consider network properties and sharing behaviour. This is reflected on the list of the most representative messages for each community (Table 4), which are usually related to few subjects. Accordingly, by aggregating information about which communities interacted with some object (a tweet, for instance) to other techniques (e.g., natural language processing), it may be possible to gather new knowledge about such object that are not evident in it (as social or geographic contexts). In an extended perspective of this result, one can wonder if it might be possible to infer significant elements about the nature of a process happening on a social network even without access to the content traveling through such network, if there is information available about its structure and dynamics.

Ethical implications of the power of such methods for detecting users specific interests, without the direct access to their personal information, should be considered. If, on one hand, this knowledge can be used in order to improve the performance of useful systems, as machine learning algorithms, on the other hand it may also incur in risks to privacy and security. This discussion is not conducted in details here, but should not be forgotten.

The communities observed had elements of cohesion on their general behaviors, emphasizing or even repressing the spread of certain types of content. An interesting conflict between individual autonomy and collective behavior seems to be part of information diffusion processes that take place on OSNs self-organized in communities. The community specialization in topics of interest is also evidenced. On a context of proliferation of many different subjects, the limitation of the scope of themes discussed within a community can be an efficient strategy for individuals to deal with information overload.

Further steps of the research include a deeper analysis of a database including more messages from multiple sources, using text-mining to define messages subjects and comparing such classification to results obtained by *tf-idf.* In future studies, the homophily of sharing behavior in online social networks can be subject of a deeper analysis, developing new methods to try to determine how much of it is due to (1) preference of individuals to establish new social connections to similar peers; (2) social influence; (3) indirect homophily, which occurs due to the existence of homophily of another trait (e.g., if two individuals usually access Twitter in the same hours of the day, despite of their connections they will be more likely to read the same news and, thus, retweet it). An even deeper analysis can be made about social influence, in order to be able to divide it into its reactive part – an individual exhibits a sharing behavior in favor of some subject because his/her community publishes more about such subject, not because of an inner preference – and its cognitive part – by observing his/her peers' behaviors, an individual shapes his/her preferences according to those practiced by his/her peers.

6. REFERENCES

[1] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct. 1999.

[2] J. Bollen, B. Goncalves, G. Ruan, and H. Mao. Happiness is assortative in online social networks. *Artificial life*, 17(3):237–251, Jan. 2011.

[3] J. Borge-Holthoefer, R. Baños, S. González-Bailón, and Y. Moreno. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, 1(1):3–24, 2013.

[4] M. Dillon. Introduction to modern information retrieval. *Information Processing & Management*, 19(6):402–403, 1983.

[5] S. Goel, D. Watts, and D. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*, volume 1, page 623, New York, New York, USA, 2012. ACM Press.

[6] A. Halu, K. Zhao, A. Baronchelli, and G. Bianconi. Connect and win: The role of social networks in political elections. *EPL (Europhysics Letters)*, 102(1):16002, Apr. 2013.

[7] D. Kurka, A. Godoy, and F. Von Zuben. Online social network analysis: A survey of research applications in computer science. *arXiv:0707.3168 [cs.SI]*, 2015.

[8] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, page 591, New York, New York, USA, 2010. ACM Press.

[9] T. Lansdall-Welfare, V. Lampos, and N. Cristianini. Effects of the recession on public mood in the UK. In *Proceedings of the 21st International Conference Companion on World Wide Web (WWW '12)*, pages 1221–1226. ACM, 2012.

[10] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1):5, May 2007.

[11] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[12] M. Mitchell. *Complexity:A Guided Tour.* Oxford University Press, 2009.

[13] M. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[14] D. Romero, C. Tan, and J. Ugander. On the interplay between social and topical structure. *arXiv:1112.1115 [cs.SI]*, 2011.

[15] M. Rosvall and C. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[16] M. Salathé, D. Q. Vu, S. Khandelwal, and D. R. Hunter. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Science*, 2(1):4, Apr. 2013.

[17] A. Sarcevic, L. Palen, J. White, K. Starbird, M. Bagdouri, and K. Anderson. "beacons of hope" in decentralized coordination. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*, New York, New York, USA, 2012. ACM, ACM Press.

[18] S. Stieglitz and L. Dang-Xuan. Political communication and influence through microblogging – an empirical analysis of sentiment in Twitter messages and retweet behavior. In *2012 45th Hawaii International Conference on System Science (HICSS)*, pages 3500–3509, 2012.

[19] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media*, 2010.

[20] D. Watts and P. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458, 2007.

[21] F. Wu, B. a. Huberman, L. a. Adamic, and J. R. Tyler. Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337(1-2):327–335, June 2004.

[22] S. Wu, J. M. Hofman, W. a. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*, page 705, New York, New York, USA, 2011. ACM Press.