Twitter Trend Topics and Their Users – an Interpretation as a Complex Network

Diogo Pacheco[1] and Ronaldo Menezes[1]

[1] BioComplex Lab, Florida Institute of Technology (FIT), United States
dpacheco2013@my.fit.edu, rmenezes@fit.edu

Abstract. Understanding trend topics generation in terms of who is more relevant or when permanent trends are more likely to happen, could be a competitive edge over social media. This paper investigates the relation between trends and users using Twitter data from trend topics from Brazil. The results suggest that a small group of users can be addressed as the more influential ones, and that target days of week and / or hours of a day can increase potential hashtags to become trend.

Keywords: Twitter Tracking, Trend Topics, Relevant Users, Social Network Analysis

1 Introduction

Until the few last years, one could simple remember trends in terms of clothes, or fashion. At the present time, the term trend itself, has become popular – a trend. Such popularity has, sometimes, paradoxical meanings. While a trend line is something more stable than the actual signal it represents, other systems may refer to trends as instantaneous most popular results. In this way, one can find trend topics, best sellers, or most listened to. Beyond any philosophical discussion [1] about the precedence or merit of being a trend, the fact is that trends are useful marketing tools and can be profitable [2-3]. Understanding how / when they are created or who creates them (if it is not a collective emergent process) would create a competitive edge.

Twitter is a micro blog application as a social network site where users can post public messages (tweets), limited to 140 characters each, and messages related to any subject. As they are posted, they can be replied to, liked, or even re-tweeted. Although tweets are public, the spread of information is dependent on how the social network is structured – in terms of the relations "following" and "follower". For instance, when a popular user (one with many followers) tweets something, his/her message is automatically delivered to all his/her followers. On the other hand, a message of a non-popular user would be reached only with specific searches. Twitter also detects and publishes its "Trend Topics", commonly referred to in this paper as trends. As the name suggests, they are related with the most popular subjects in the moment, "the breaking news" (according to Twitter internal analysis of all posted tweets).

In this paper, the relationships between trend topics and their users were investigated from the perspective of 2-weeks of data from Brazilian trends. The results show that a small-representative group of users can be identified among others. In addition, statistical analyses demonstrate that there are preferential days / hours to create shorter or longer trends.

2 Related Work

Social media applications (Twitter, Facebook, YouTube, etc.) have been used by several knowledge areas to study and explain real world phenomena [4-6]. The availability of API or techniques for web crawling made huge datasets accessible.

Much research has been done locating the most important users (central, influential, popular, among others) [7-9], while others focused on trends' formation process. Barabasi explained shocks and bursts and how the structure of a network can contribute to trending [10]. The study performed by KISSMetrics aimed to explain the time of the day in which one is more likely to be re-tweeted [11]. Zubiaga *et al* categorized trends by the happenings that raise them [12].

Inspired by those studies, this paper confronts users' centrality for trend creation and addresses different classification according to recurrence.

3 Building a Twitter Network

There are two common approaches when one intends to build networks: starting by gathering data or by modeling. In this work, the first step was to collect data from Twitter and then two different network models were proposed.

3.1 Collecting Data

A Twitter tracking application was developed in Python 2.7 using the Python Twitter Tools API [13]. Between September 19th and October 2nd of 2013, at every 15 minutes, the trend topics from Brazil were requested using the API method *trends/place* (with ID parameter WOEID[14]), receiving at each iteration 10 new or recurrent trends.

Two threads were created for each new trend: one responsible for retrieving related tweets from the past (i.e. tweets posted before the trend became trend), using the method *search/tweets*, and constrained to Portuguese language in a six-hours time window, and another responsible for listening to the future (i.e. tweets posted after the trend has been created) using the streaming method *statuses/filter*. In contrast, when a trend was recurrent, its counter would be incremented and a thread for listening for the following in the future would be started, if it had not been already running. If a trend had not been listed in the top rank for a consecutive-45-minutes then its forward thread would has stopped.

A MySQL database was used to store trends, users, and tweets data. In **Table 1** the parameters' configuration are detailed. **Table 2** shows the volume of collected data for this work.

Table 1. Configuration set-up for twitter tracking.

Parameter	Value
Start Date	9/19/2013
Finish Date	10/2/2013
Backward Window Time	6h
Forward Window Time	45min
Constraints (methods)	
trends/place - id	23424768
search/tweets - lang	pt

Table 2. Number of tracked trends, tweets, and users, counted before and after trend creation.

	Before Trend	After Trend	Total
Trends			1,760
Tweets	3,491,163	3,066,883	6,558,046
Users	775,684	1,398,988	1,921,077

The recurrence of a trend was measured by its counter attribute. Although all of them are trend topic, the counter distribution is not normal. **Fig. 1** shows the cumulative absolute distribution of trends by their counter in a log-log chart, i.e. all 1760 trends have counter greater than or equal to 1, but only one has it greater than 288. The number of tweets posted by users also does not fit in a normal distribution. While most users tweets only about one topic, a few others may tweet over a hundred trends (see **Fig. 2**).

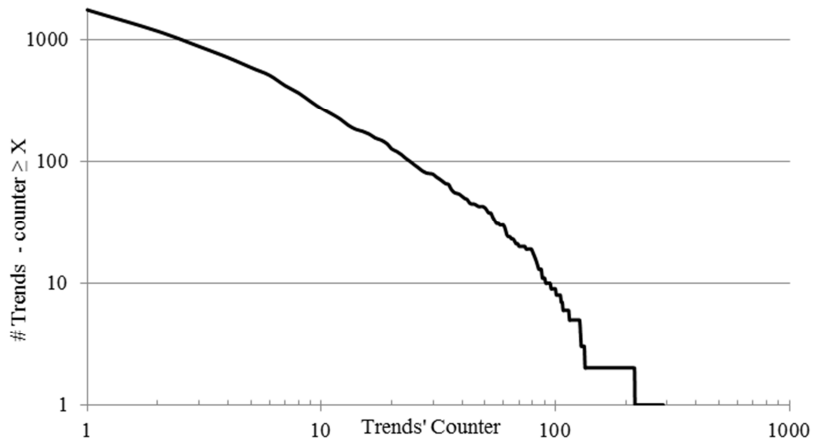

Fig. 1. Distribution of 1760 trends according to their counter, varying from 1 to 289

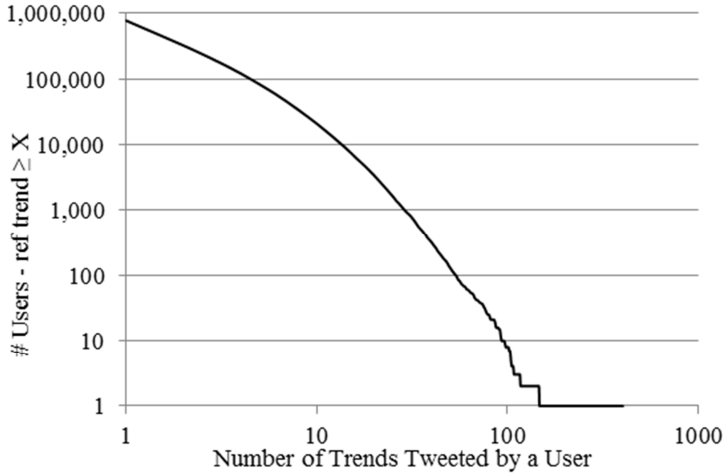

Fig. 2. Distribution of number of trends tweeted by a user, varying from 1 to 402

3.2 User-Trend Bipartite Network

A natural model emerges from the tweets' relationship between users and trends, i.e. creating a bipartite network where nodes are users and trends, edges represent tweets from a user mentioning a trend, and the weight is the number of tweets. However, the recommendation is to project the network due to little information that can be revealed from this structure. Although a user projection network suggests being more meaningful than the projection over trends, that network could not be addressed in this paper because of high network density and elevated time-consumption.

3.3 User-User Directed Network

Instead of attempting to understand the subjects' pattern among trend topics, we were interested in discovering who is responsible for their creation. In the proposed model, when a tweet is re-tweeted or replied to, a direct link from the second user to the one who started the communication would be created. The relation's weight is the number of times that a user replied or re-tweeted another user. Note that in this approach never-mentioned tweets will not affect the network structure. Moreover, the relation formed by users and their tweets is not transitive due to Twitter's structure. For instance, if a user B re-tweets a user A, B will propagate that information to his followers. Then, if one of those followers C re-tweets B, in fact, they will re-tweet A. Hence, the network would have only the edge from C to A.

Only data collected before trends had been created were used to build the user-user directed network. In **Fig. 3**, three layers of visualization of this network are displayed. To allow the identification of key users on trends formation, four different metrics were suggested to be applied using Gephi0.8.2[15]:

- Weighted-In-Degree (WID) – reveals the ones who generate relevant information, considering that the propagation is proportional to the relevancy;
- Betweeness (BTW) – shows ones who promote several trend topics (spread), those who can generate relevant content through different trends;
- Page-Rank (PRK) – show the ones who are ideally the trendsetters, or from whose ideas the information becomes relevant;
- Weighted-Out-Degree (WOD) – suggests ones who intensively promote a trend topic.

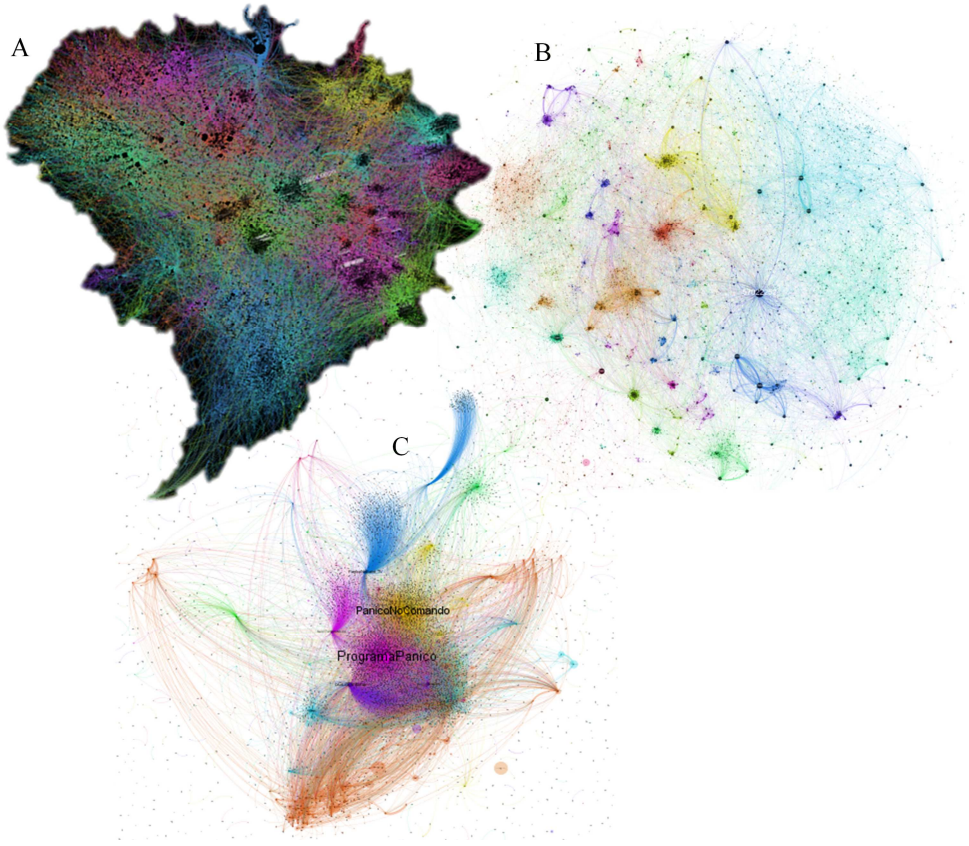

Fig. 3. Tree layers of visualization from the user-user network: (A) the entire network, (B) the giant-component of those with degree > 2, and (C) the network of a single trend with 15,501 nodes / 26,892 edges.

4 Experimental Results

4.1 User-Trend Bipartite Projections

Before making a projection, the user-trend bipartite network was created. Some metrics were calculated, but aiming to correlate subjects from trends, this model focused on community detection. The modularity algorithm identified 33 communities with a resolution parameter as 0.73. The bipartite network is shown on **Fig. 4**, but due to size and resolution, one can argues that it is not bipartite. In fact, what looks like nodes are conglomerates of trends and theirs users. A particular community was highlighted and then filtered, revealing the related trend topics.

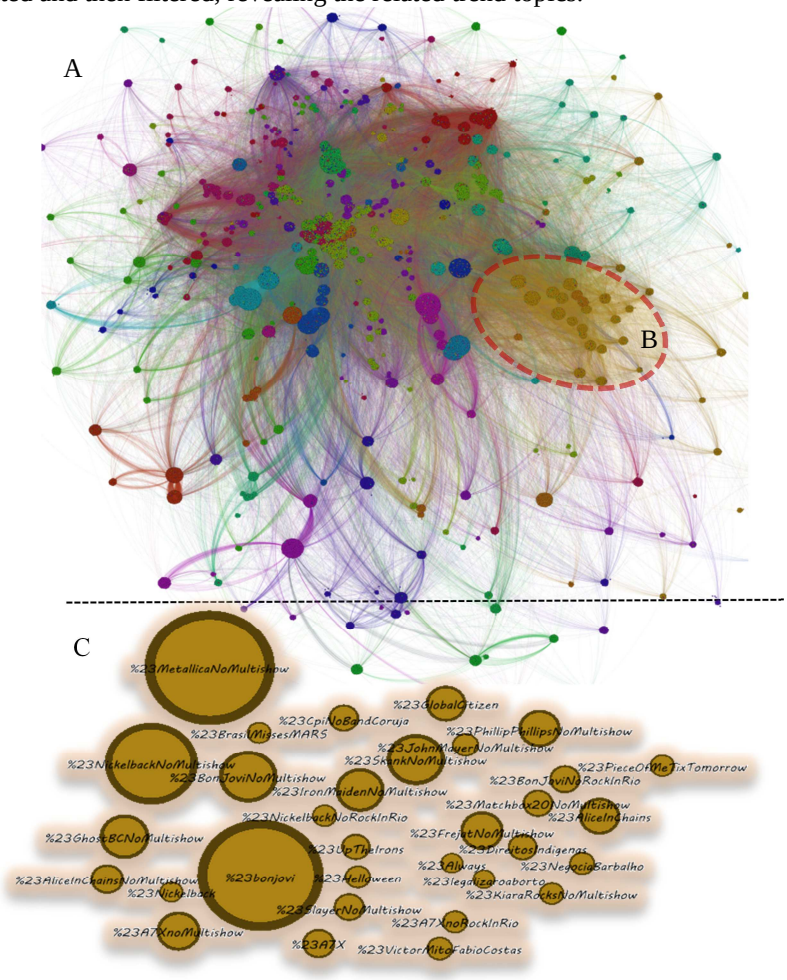

Fig. 4. Network bipartite of users and trends where nodes' size is proportional to in-degree (A). A known community formed by trends related to Rock-In-Rio Festival (B). Zoom-in in the community filtering and showing only trend nodes (C).

The application of a modularity algorithm tended to find three major communities on trends projection network even when the resolution parameter was significantly changed. **Fig. 5** shows the trends projection network divided into numbers of communities similar to the bipartite network community classification. Trends of the highlighted community on **Fig. 4B** were classified inside one of those three major groups.

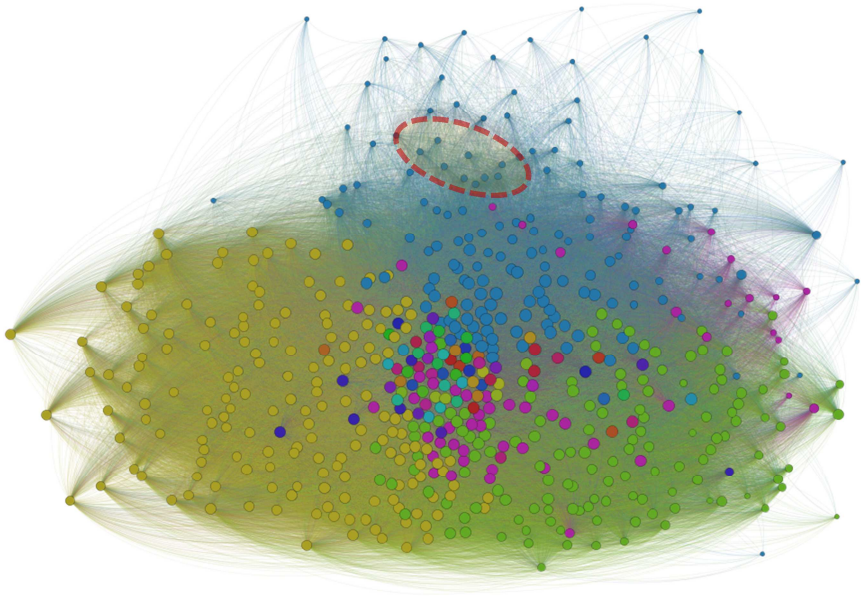

Fig. 5. Network of trends (39 colored communities), generated through a projection from a user-trend network

4.2 User-User Directed Network

To determine the most relevant users according to definition presented in section 3.3, a preliminary list was created with the top-20 users of each metric. The final list, presented in **Table 3**, contain the top-3 users of each metric plus those ones who were listed in more than one list. At last, 26 users have been chosen.

To check how representative the selected 26-users group is for the entire network formed by the Brazilian trend topics, three different sized groups (26, 100, and 1000 users) were randomly created. **Table 4** details trends coverage and the direct traffic generated by each group where the random results are the average of 10 runs. Note that 1,000 users were directly responsible for less than 2% of all tweets, i.e. the number of tweets posted plus the number of re-tweets over those tweets. In contrast, the top-26 users' contribution extends 3.5% of all traffic. In addition, these users tweeted about 46% of all trends.

Table 3. Top-26 relevant Twitter users from user-user direct network according to weighted-in-degree (WID), weighted-out-degree (WOD), page-rank (PRK), and betweenness (BTW). Friends and Followers are Twitter attributes and Trends is the number of different topics tweeted by each user.

ScreenName	Friends	Followers	Trends	WID	WOD	PRK	BTW
rockinrio	519	1551709	70	4		3	9
multishow	96165	896993	91	9		7	2
felixpassiva	37	95048	43	10		9	11
jooseanee	945	104924	77	8		11	10
luscaspfvr	333	79896	53	13		15	5
SignosFodas	18	1394924	77	1		1	
justinbieber	121293	46915644	10	2		4	
MTVBrasil	2055	1234162	17	6		5	
programapanico	1967	7971179	28	3		6	
bandmichael	194	51054	16	7		8	
PanicoNoComando	99	180697	49	5		10	
EuNicoleBahls	1583	259825	11			12	7
oguisantana	872	439898	10	15		13	
sophiaabrahao	186	1373903	18	11		14	
cassiosuruba	1793	1329	7			16	6
NosTrendsBrasil	241	79365	76	17		17	
cauemoura	521	396828	19	16		18	
Rhiitler	195	44594	30	14		19	
instagranzin	129071	508517	66	12		20	
_RioGrandeDoSul	1519	37307	15	18			4
SignosOGuru	78	122145	1			2	
trendinaliaBR	16	2852	402				1
atanair	10849	53242	33				3
feelsanitta	184	2553	25		2		
PerfeicaoAnitta	301	1328	37		1		
comvoceanitta	251	1182	22		3		

Table 4. Number of trends, users, tweets and re-tweets (RT) for networks of best 26-users, and random 26, 100 and 1000 users, confronted to the complete one.

	Users	Tweets	RT	Tweet+RT		Trends	
Best 26 users	0.003%	5461	117565	123026	3.52%	757	43%
Random 26 users	0.003%	1222	3827	5049	0.14%	177	10%
Random 100 users	0.013%	3512	998	4510	0.13%	393	22%
Random 1000 users	0.129%	38756	21675	60431	1.73%	1180	67%
All	775684	3491163	3491163	3491163		1760	

4.3 Trend Analysis

Besides network relations, analyzing the raw data can also reveal significant information. **Fig. 6A** shows the tweets distribution along the week and that there were more tweets posted on Friday, Saturday, and Wednesday in Brazil in the analyzed period. In contrast, **Fig. 6B** shows topic trends creation in days of the week. In this case, they were more likely to be formed on Sundays, Saturdays, and Fridays. While a decreasing pattern from Sunday to Thursday and a growing pattern from Friday to Sunday is clearly observed from trends distribution (**Fig. 6B**), any apparent behavior is detected from distribution of tweets (**Fig. 6A)**.

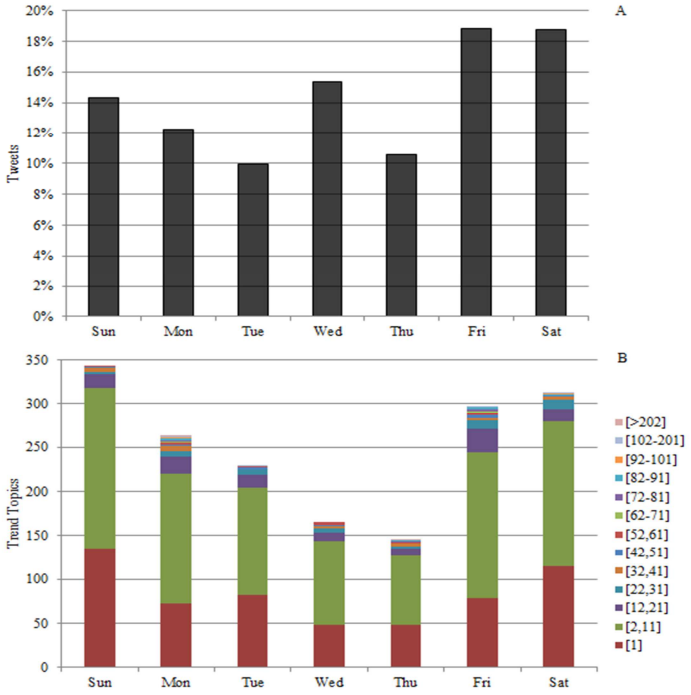

Fig. 6. (A) Percentage of tweets per day of week. (B) Number of new trends created per day of week and colored according to theirs counter.

Since there was no strict relation between the number of tweets and the number of trend topics, a deeper analysis over trends formation was performed. As showed in **Fig. 6B**, trends were classified according to their relevance, in other words, higher the counter more permanent a trend was on the top, i.e., counter=1 means that a trend was identified in the Twitter top rank once, during the 2-week-period analyzed. **Fig. 7** shows six groups of trends and their percentage of creation during days of the week or according to hours of the day[1].

[1] Charts display hours according to (UTC-05:00) Eastern Time.

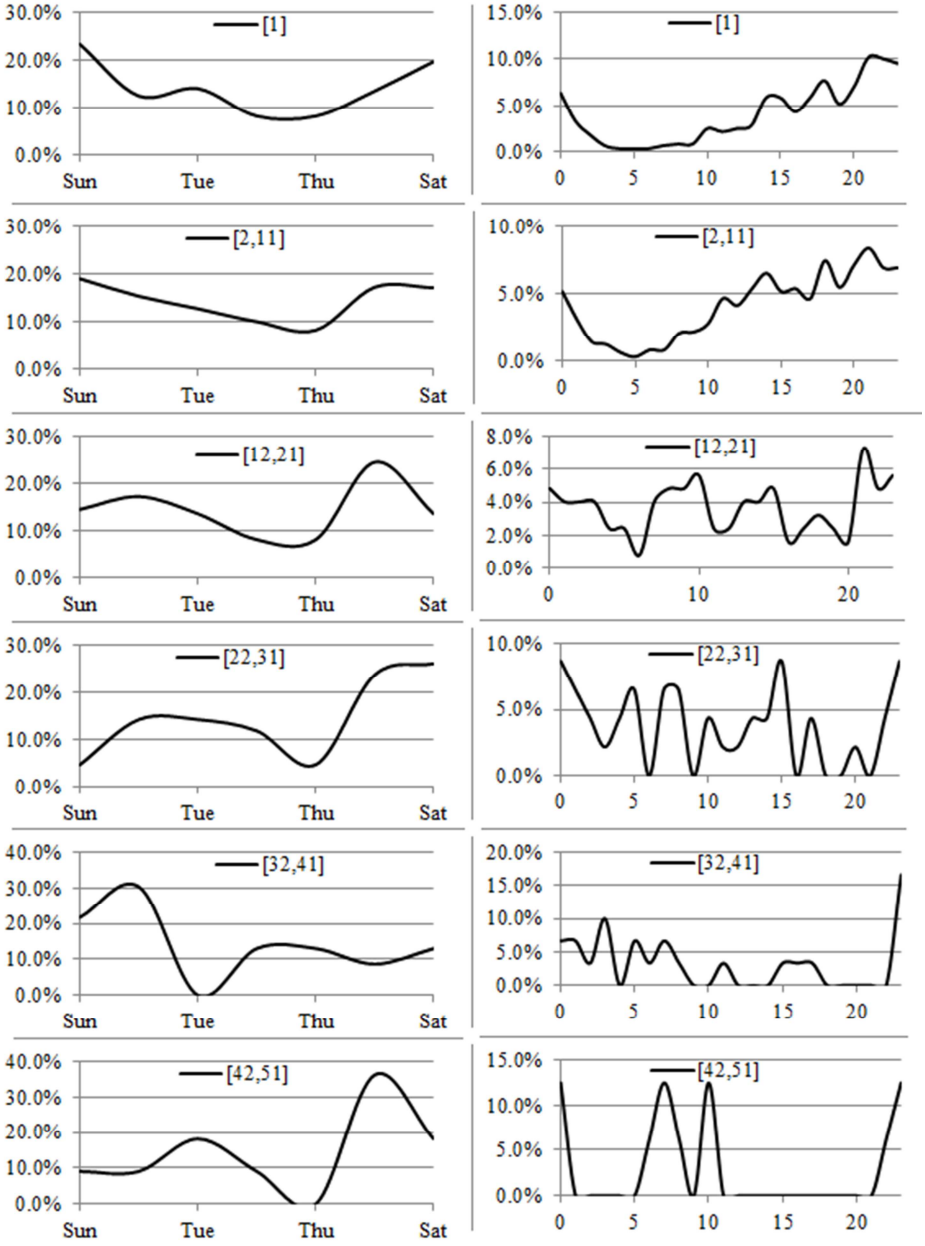

Fig. 7. Trends' probability of occurrence during days of week (left-side) and hour in a day (right-side). Treads are grouped by their counter.

5 Conclusions

This work collected and analyzed tweets from Brazilian trend topics between September 19th and October 2nd of 2013. The objectives were to discover important features related to trends formation, such as (but not limited):

- Who creates trends?
- What are trends about?
- When are trends created?

Two models were proposed to try to reach those objectives. The first one (and more natural) created a bipartite network with users and trends as nodes. Nevertheless, the projected network of users and its analysis has been shown to be prohibitive due to available resources and time constraints for this work. While the projection of trends was realized, it does not aggregate relevant information. In fact, the projection seemed to lose meaningful information about communities when compared to the original bipartite network.

The second model generated a network based on tweets popularity, where links were directly created between users if one had replied or had re-tweeted the other. Even though this model could not reveal features by simple visualization, by network metrics analysis it was possible to identify a small group of relevant users in trends formation. Those could be targeted by marketing companies as promoters, for example. An interesting finding in this model was that the lack of transitivity in the Twitter relationship made the weighted-in-degree (WID) metric very significant in finding the best users. As Twitter analysis can handle dense user networks, WID can reveal goods estimation being much less time-consuming than other metrics, such as, page-rank. In contrast, the weighted-out-degree seemed to be irrelevant suggesting that posting thousands of tweets is not enough. To be successful in trend generation, one's information must be propagated by others, and it is totally dependent of one's network structure.

Finally, the statistical analysis on trends data indicates that different kinds of trends (according to their relevancy) tend to be created on different days and hour. Also, the likelihood of a trend topic formation may not be inferred by the volume of tweets.

These findings should be verified in other places besides Brazil in order to be generalized as a common behavior for trend formation. As the present study has considered a short time period study, those user are best restricted to that period. A longer-period study could expose the permanent-best users.

References

1. Theosophy (September 1939). "Ancient Landmarks: Plato and Aristotle". Theosophy 27 (11): 483–491. http://www.blavatsky.net/magazine/theosophy/ww/additional /ancientlandmarks /PlatoAndAristotle.html
2. Ferguson, Rick. "Word of mouth and viral marketing: taking the temperature of the hottest trends in marketing." Journal of Consumer Marketing 25, no. 3 (2008): 179-182.
3. Celente, Gerald. *Trends 2000: How to Prepare for and Profit from the Changes of the 21st Century*. Hachette Digital, Inc., 2009.

4. Davidov, Dmitry, Oren Tsur, and Ari Rappoport. "Enhanced sentiment learning using twitter hashtags and smileys." In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 241-249. Association for Computational Linguistics, 2010.
5. Dwyer, Catherine, Starr Roxanne Hiltz, and Katia Passerini. "Trust and Privacy Concern Within Social Networking Sites: A Comparison of Facebook and MySpace." In *AMCIS*, p. 339. 2007.
6. Paolillo, John C. "Structure and network in the YouTube core." In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pp. 156-156. IEEE, 2008.
7. Trusov, Michael, Anand Bodapati, and Randolph E. Bucklin. "Determining influential users in internet social networks." *Available at SSRN 1479689* (2009).
8. Newman, Mark EJ. "Scientific collaboration networks. I. Network construction and fundamental results." *Physical review E* 64, no. 1 (2001): 016131.
9. Zhang, Yu, Zhaoqing Wang, and Chaolun Xia. "Identifying key users for targeted marketing by mining online social network." In *Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on*, pp. 644-649. IEEE, 2010.
10. Barabasi, Albert-Laszlo. "The origin of bursts and heavy tails in human dynamics." *Nature* 435, no. 7039 (2005): 207-211.
11. The Best Time Of Day To Tweet (To Get The Most RTs), KISSMetrics, http://www.bitrebels.com/social/the-best-time-of-day-to-tweet-to-get-the-most-rts/
12. Zubiaga, Arkaitz, Damiano Spina, Víctor Fresno, and Raquel Martínez. "Classifying trending topics: a typology of conversation triggers on twitter." In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 2461-2464. ACM, 2011.
13. Python Twitter Tools, http://mike.verdone.ca/twitter/
14. Yahoo, Where On Earth ID - GeoPlanet, http://developer.yahoo.com/geo/geoplanet/
15. Gephi, an open source graph visualization and manipulation software, https://gephi.org