Twitter Community Detection Using Principles of Dynamic Social Impact Theory

Benjamin Mitchell[1], Partha Mukherjee[2], and Youakim Badr[3]

[1, 2,3] The Pennsylvania State University, Great Valley, Malvern, PA-19355

Abstract. With the explosion of online communication, understanding how public perception is shaped by information flow across social networks is of increasing importance to researchers and corporate interests alike. This research provides analysts the framework they need for understanding the relationships between influential subcultures across the US, at a specific snapshot in time. In this study we intend to use Dynamic Social Impact Theory (DSIT) as a guide for understanding community clustering methods, using large quantities of social media data. Our objective is to develop a framework for modeling networks of Twitter users based on cultural attitudes gleaned from their online behavior. The outcome of our study finds remarkably interconnected networks of people containing clearly delineated subcultures which are conducive to reinforcing and amplifying internally held opinions and worldviews.

Keywords: Social Networks, Dynamic Social Impact Theory, Community Detection.

1 Introduction

Dynamic Social Impact Theory (DSIT) [1] is based on the idea that culture emerges from and is shaped by certain individuals who exert social influence. The cultural groups that form as a result of these individuals' influence are constantly evolving, continually recalibrating themselves as a result of communication: "Culture is not a static concept. Elements of culture change over time, as does the popularity of different cultural worldviews" [2].

Furthermore, DSIT states that these subcultures are reorganizing themselves along four basic principles: clustering, correlation, consolidation, and continuing diversity, which are described as phenomena of culture: 1) clustering, or spatial diversity of elements, 2) correlations between elements because of clustering, 3) consolidation defined by temporal reduction in diversity, and 4) continuing diversity [3].

Historically, subcultures (notable examples include hippies, bohemians, and various political movements) have all followed these self-organizing principles: self-clustering leads individuals to surround themselves with people whose beliefs reflect their own, then correlation occurs as opinions solidify from those with local social influence, promoting the spread of similar ideas or ways of thinking, until eventually these opinions become uniformly adopted during consolidation, where all group members share a similar outlook and worldview. These intra-cluster dynamics have led many researchers [4-6] to predict the inability of majority groups to

dominate entire networks, because groups holding minority opinions are insulated from dissent by their like-minded membership.

In our study a framework is developed to identify the subcultures which form as a result of clustering. Our results will be presented as social network graphs, depicting the internal structure of each subculture as well as the relationships between respective subcultures. Our study will introduce a new method of identifying connections between Twitter users, and demonstrate that modern community detection methods are based on underlying DSIT assumptions.

2 Literature Review

With the advent of social media there was a paradigm shift in the landscape of interaction. Social media is a spectrum of multiple online social network platforms such as Facebook, Twitter, Weibo just to mention a few [7]. People could indulge upon information seeking and searching behavior on diverse aspects such as social, political, business, entertainment by sharing different types of posts (e.g., textual, webpage, multimedia etc.). Social media in a way establishes new forms of communication channels where people could respond both instantly and asynchronously [8]. Social media is a broad umbrella of new online communication channels. It has enabled people all over the world to interact and share information with each other via user generated content.

A study by Latané and Bourgeois [9] provides empirical evidence for DSIT and opinion clustering. In the study of a 14-story student dormitory, the authors tested cultural markers ranging from drug and alcohol usage to food and clothing preferences by means of measuring consensus of opinions. The authors found that opinions are significantly more alike for students residing on the same floor compared to those on other floors[2]. Follow-up studies also confirmed that this opinion/attitude clustering was caused by communication [10-13]. It was observed that clustering index increased on 13 out of the 15 types of opinions/attitudes measured with a significantly large overall increase [2]. Clustering identifies the variation in social opinions and attitudes which are different in different regions. This difference in cultures identified by clustering can be used to predict DSIT [3]. Ritcher and Kruglanski [14] stated that according to DSIT, people staying in a region recursively impact others by means of communication. Such a spatial communication process generates subcultures in terms of harmony in attitudes, behavior and values [15]. Though self-organized cultures indirectly assist the concept delineated by DSIT, it does not provide the support for the emergence of difference by means of everyday interaction. Bowen and Bourgeois [11] concluded that, “subcultures of beliefs and behaviors form reliably in a relatively short period of time on a variety of intellective and judgmental tasks” [4]. We plan to examine how these subcultures relate to each other, by using social media data as a basis for modeling communications between cultural groups, where cultural groups are identified using the principles described in DSIT [1].

3 Data Collection

We collected large volumes of tweets on a daily basis throughout the month of September, 2019. Since Henrique's algorithm [16] was invoked using a command-line interface, we were able to automate data collection for each of the 50 states in the US using a mixture of the Henrique package built using Python and Cron on Linux. The data collection script is executed daily at 9:00PM EST (as scheduled by the cron server), generating a shell script containing one line for every state in the US. This shell script then invokes separate daemon instances of download for each state, and each state's output is saved to 50 separate CSV flat files (one for each state) from where they can be picked up by the Apache Spark cluster and merged.

By such, our data collection process allows us to bypass the rate limit of the default Twitter API among other limitations, as well as tag each tweet's geographic location by state. Our goal was to obtain a 1% random sample of all the tweets in the US, so our queries did not involve a targeted keyword search. At the end of the month, we obtained 32 million tweets. The contents of these tweets could then be used to create cohesive subgroups of subcultures which can be plotted using social network graphs.

Social network analysis (SNA) is used to visualize the strength of their relationships, defined by intensity and frequency of communication between and within groups. The objective is to use clustering to create cohesive subgroups, which could then be plotted using social network graphs using Python's NetworkX, and Gephi packages to analyze relationships between subgroups.

Unstructured text data is prepared for further analysis by means of data cleaning. Since opinion and attitude clustering focus on extracting general themes and ideas from individual tweets and users, it is important to leave the tweets largely intact. Since hashtags often contain words that have been concatenated together (for example, "LittleArmenia"), it is important to separate these into individual words. This is accomplished with Python's "Wordninja" package (i.e., "LittleArmenia" becomes "Little Armenia"). The last step of the data cleaning process involves removing text that does not represent English words. To filter out non-English words, we reference our text data against an English dictionary [17] that is used by users on Twitter.

4 Social Network Modeling and Visualization

Our novel method of social network modeling constructs a realistic network based on how Twitter users are actually perceived by the public, by connecting users only when they are mentioned in a similar context. This method relies on Twitter mentions ('@') exclusively, and is not only more efficient than constructing a traditional network graph from users' friends' lists, but also more accurate. It is also better suited for a DSIT study, which views connections between individuals as being rooted in shared identity. Users are connected only if they are both found within another user's Twitter mentions. Strength of user relationships will be determined by the frequency by which they appear together within another user's list of mentions. These strength values are reflected in the weight column of a weighted edge list, a subset of which is shown in Table 1, in decreasing order of weights for a 100 node user network.

Table 1. Weighted edge list of mentions.

	User_1	User_2	Weight
0	SpeakerPelosi	realDonaldTrump	2803
1	GOP	realDonaldTrump	1743
2	RepAdamSchiff	SpeakerPelosi	1648
3	POTUS	realDonaldTrump	1605
4	RepAdamSchiff	realDonaldTrump	1520
5	LindsayGrahanSC	realDonaldTrump	1453
6	JoeBiden	realDonaldTrump	1323
7	realDonaldTrump	senatemajldr	1315
8	GOP	SpeakerPelosi	1246
9	JoeBiden	SpeakerPelosi	1115

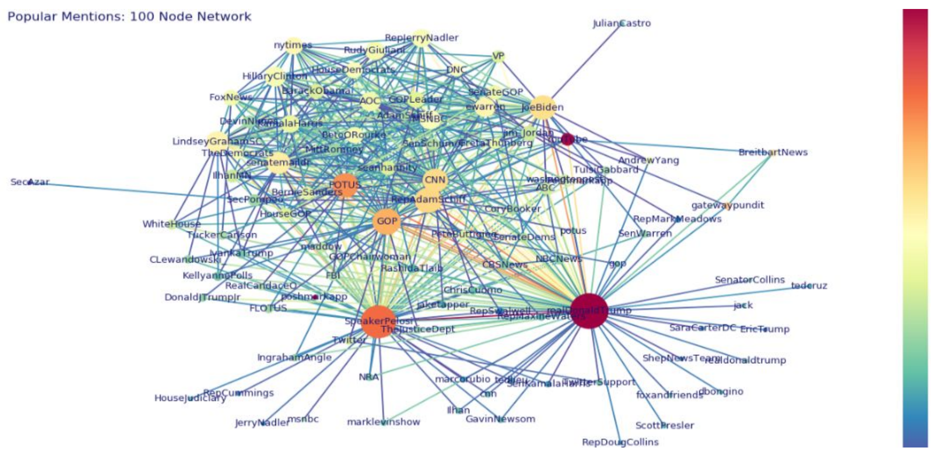

Fig. 1. Twitter Network of 100 users, with degree, user popularity, and connection strength.

The weighted edge list is based on the strength of user relationships among influential users, defined by having at least 30 tweets each, filtered in Apache Spark. Since weights are calculated on a per user basis, as opposed to per mention, duplicated mentions are removed from the mentions lists. These edge lists shown in Table 1 can then be used to connect nodes in a network. In addition to modeling degree (number of connections) represented by node size, an additional node attribute termed user popularity was added and modeled by color gradient (see Figure 1) that represents the number of mentions per user. The network in Figure 1 is created using Python's networkX package from the first 535 rows of the edge list (see Table 1) for a cleaner visualization.

Data read in from a popularity data frame as shown in Table 2 (i.e., in decreasing order of mentions per user) is used to assign a popularity attribute to each node, which was visualized with a logarithmic scale in Figure 1.

Table 2. List of mentions and corresponding counts per user

	Mentions	Counts
1	RealDonaldTrump	52926
2	poshmarkapp	51997
3	Youtube	49071
4	ebay	19853
5	SpeakerPelosi	19700
6	Etsy	16394
7	POTUS	14665
8	GOP	11483
9	BloggyMoms	10683
10	Poshmarkapp	10669

A type of community from Social Network Analysis (SNA) theory known as k-cliques seem to particularly exemplify these self-organizing principles. Intuitively, "a clique in a social network as a cohesive group of people that are tightly connected to each other (and not tightly connected to people outside the group)" [18].

As depicted in figure 1, the 4-clique composed of BreitbartNews, gatewaypundit, YouTube, realDonaldTrump, is remarkably isolated, compared to the majority of nodes in the network. The two nodes gatewaypundit and BreitbartNews both have a degree of only 3, unusual for their relatively high popularity. Their low degree indicates that it will be difficult for external information to penetrate the 4-clique, and when combined with the relatively high popularity of the 4-clique, ideas from within the clique propagate efficiently, with little influence from the outside. This behavior exemplifies both the "culture-generating" and "amplification properties" from the definition of cliques.

Self-clustering behavior was analyzed further using the Louvain method [19] for community detection. We chose the Louvain algorithm as it is a popular modularity optimization method that is fast and scalable for handling large social networks. The Louvain method detected five distinct communities in our Twitter network (100 node, 535 edge subset), generating a visualization similar to Figure 1. Each community was labeled according to its members' Semantic Web Identity (SWI) [20] i.e., a contextual categorization of entities by Internet search engines. SWI was chosen because it is a widely used method of identification, in addition to having a proven ability to infer relationships between entities. In our study, SWIs were gathered manually by a native English speaker, although this step could easily be automated for future research.

Common SWIs for each community are listed in the Description column of Table 3. It should be noted that these naming conventions are irrelevant in the context of DSIT clustering, where subcultures are defined by their members, not their labels. The descriptions did, however, allow us to better understand the internal structures of each community. The observed uniformity of SWIs across community membership substantially increased our confidence in the ability of the Louvain algorithm to categorize Twitter users, and provided a window into how relationships between community members are rooted in shared identity, consolidated over time by the third DSIT principle.

Table 3. Community detection using Louvain algorithm in a 100 node network.

Comm ID	Description	Notable Users	Total Members
0	Politician: Liberal	SpeakerPelosi, JoeBiden, AOC	36
1	Right-wing media	FoxNews, BreitbartNews, realDonaldTrump	35
2	Politicians: Conservative	GOP, LindseyGrahamSC, Jim_Jordan, VP	17
3	"Mainstream" media	CNN, nytimes, maddow	10
4	Advertising/bots	Poshmarkapp, poshmarkapp	2

Community detection is employed to illustrate the dynamics between influential subcultures. While performing community detection on the smaller 100 node Twitter network demonstrates its capability in categorizing popular users, applying it on a much larger scale allows us to identify the largest, most influential groups of users on Twitter.

A 100K edge list was used to create this larger network, consisting of 8690 nodes total. Reducing this network to its giant component (connected nodes only) substantially increased the power of the community detection algorithm. According to social network theory, a giant component constitutes a network subset in which every node is reachable from every other. The fully connected giant component was made up of 7297 nodes, or 84% of the larger Twitter network.

Networks were mapped out using a force-directed layout, a spatialization process implemented with the Force Atlas 2 algorithm [21] using open source Gephi visualization and exploration software [22] to generate the communities in Figure 2. This arranges the nodes in an intuitive fashion that allows for internal and external community structure analysis, corresponding with the central goals of this study.

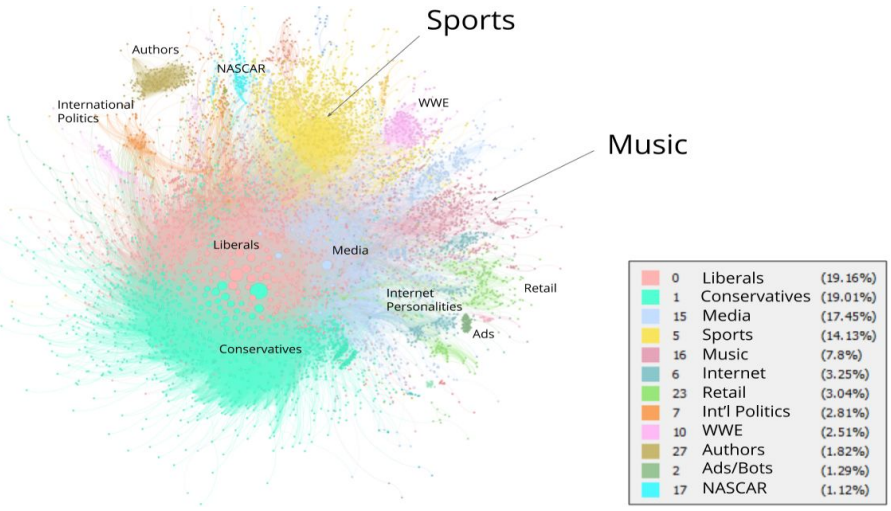

Fig. 2. Detecting Louvain communities in larger Twitter network with 7K nodes

The Louvain algorithm detected a total of 64 communities within the larger Twitter network. Of these, the 12 largest subcultures are labeled in the visualization, arranged in the legend by fraction of the network covered by each community (see Figure 2, where legend contains: community color, ID, description, network fraction). This map helps us to examine the extent to which users belonging to different communities, or subcultures of society, interact with each other. The Louvain algorithm seemed to perform well, generally categorizing individuals along the four DSIT principles, each discussed in the following section.

5 Discussions and Implications

Subcultures identified via community detection algorithms are shown here in more detail to demonstrate their strong internal uniformity. The three largest communities (see Figure 3) located near the center of the network resemble the groups analyzed in the smaller 100 node network i.e., liberals, conservatives and media. The fourth most influential group was largely composed of sports personalities (see Figure 4).

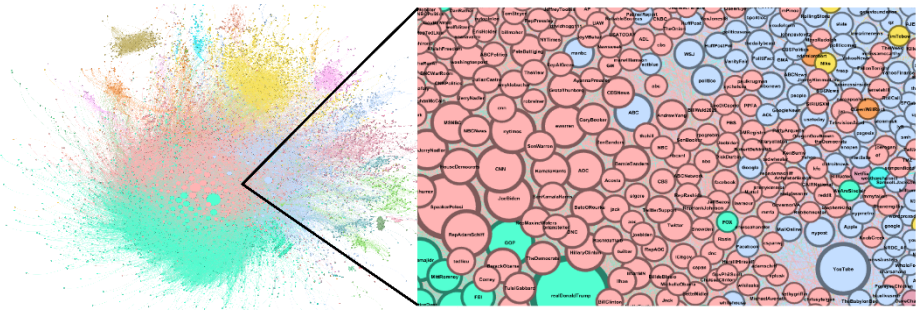

Fig. 3. Three largest Louvain communities color coded in cyan, pink and sky blue.

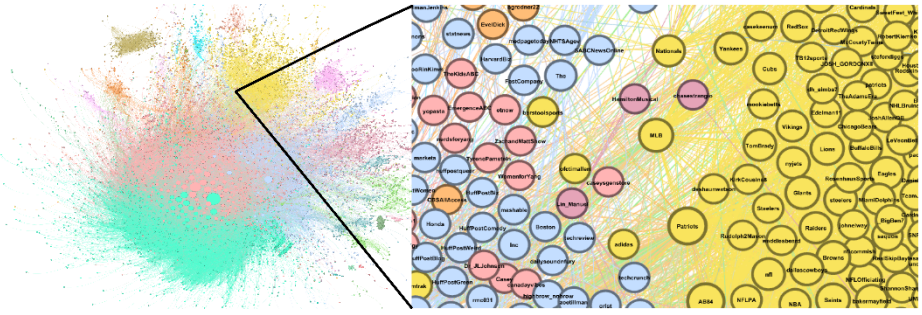

Fig. 4. Sports community detected by Louvain algorithm, shown in gold.

Musicians made up the fifth most popular subculture (see Figure 5). Most subcultures were highly interconnected, as evidenced by the highly connected nature of this Twitter network, where the giant component accounts for 84% of the network.

The internal hub-spoke structure of these communities shares many features with a method of constructing networks termed ego networks [23]. Community members are clustered around central high degree nodes. This internal hierarchical structure backs up the first DSIT assertion that social influence of certain individuals' shapes (sub)cultures. Externally, communities also display local social influence over one another. Nearby communities have much more in common than those on opposing ends of the network.

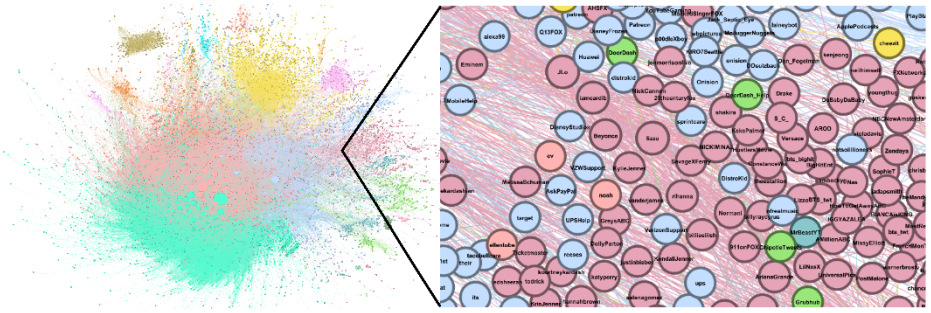

Fig. 5. Music community detected by Louvain algorithm, shown in purple.

The relationships between communities on opposing ends of the network recalls the concept of isolated cliques discussed earlier. The cliques identified in the smaller networks exemplify the second DSIT principle of correlation leading to the third principle of consolidation. Although these cliques were the most extreme example of this phenomenon we identified, the internal structure of all communities within the network is indicative of a tendency to reinforce and amplify opinions and worldviews, through correlation and consolidation, respectively. Organized around several key players, or high-degree nodes, each subculture is shaped by such individuals exerting local social influence.

We also made observations on the fourth DSIT principle, continuing diversity, which states that individual communities never completely take over the network. No single community is clearly dominant over our larger Twitter network, in fact the three largest communities (see Figure 3) cover nearly equal fractions (19.16%, 19.01%, and 17.45%). This observation reflects Harton and Bourgeois's assertion [2] that no singular culture is likely to ever dominate the world, assuming nonlinearity, which Latané Nowak's catastrophe theory of attitudes (CTA) [6] recognizes as a necessary element of self-organized networks. According to CTA, attitudes are less susceptible to change, or nonlinear, when people become more involved in issues. We recommend repeating our study several times over the course of a year to observe the fourth DSIT principle over a longer interval. The flow of information throughout the network is dictated by these internal and external structures.

The ability to gather, at a glance, relative proximity, influence, and power of any given individual puts tremendous power in the hands of marketers, intelligence experts, political organizations, or anyone wishing to better understand efficient means of information diffusion. By studying our social mapping visualization, individuals have the ability to locate specific groups, communities, or cultures they wish to target with information, and the individuals within them that may help spread

it the fastest. If the competitive arena of the future is said to be based on information, then the results of our study could very well be its map.

6 Conclusion

In our research, we generated user communication networks from Twitter mentions. The clustering behavior of the communities detected within each network showed how the Louvain algorithm operates on the principles of Dynamic Social Impact Theory.

We recognize that this research represents a novel contribution to its field. Although the relative popularities of many of the individuals depicted in our Twitter network is common knowledge, the mentions-based social mapping technique presented in our research is the first of its kind, to the best of our knowledge. Our social network reveals diverse, highly connected, and clearly defined American subcultures, each clustered around influential central actors who encourage the uniform adoption of worldviews which correlate, consolidate, and persist over time.

To better understand how communities on opposing ends of the network are related, we will run this experiment repeatedly to explore network dynamics. We acknowledge the limitations of examining DSIT over a month-long snapshot in time, and recommend adding a significant temporal aspect to the study, which would provide valuable insight into how these subcultures originated and developed over time. This would help further our understanding of the DSIT principles of consolidation and continuing diversity, allowing us to pinpoint worldviews which have exploded in portions of the population, relative to others which have faded away. We will automate the detection of SWIs that will help expedite community labeling.

In an age where political outcomes, corporate interests, and bottom lines are increasingly tied to public perception and the flow of information, these concepts are now more relevant than ever before, and our research provides analysts the framework they will need to more concisely understand and model social networks.

References

1. Nowak, A., Szamrej, J., Latané, B.: From private attitude to public opinion: A dynamic theory of social impact. Psychological Review. **97**(3), pp. 362-376. (1990).
2. Harton, H. C., Bourgeois, M. J.: Cultural elements emerge from dynamic social impact. The Psychological Foundations of Culture, Editors, Schaller, M. and Crandall, C.S. page 60. Lawrence Erlbaum Associates. Mahwah, New Jersey. (2003).
3. Harton, H. C., Bullock, M.: Dynamic social impact: A theory of the origins and evolution of culture. Social and Personality Psychology Compass. **1**(1), pp. 521-540. (2007).
4. Schaller, M. and Crandall, C.S. The psychological foundations of culture. Lawrence Erlbaum Associates, London, Mahwah, New Jersey. (2004).
5. Latané, B., Nowak, A.: Self-organizing social systems: Necessary and sufficient conditions for the emergence of clustering, consolidation, and continuing diversity. Progress in Communication Science. 13, pp. 43-74. New York: Ablex. (1997).
6. Latané, B., Nowak, A.: Attitudes as catastrophes: From dimensions to categories with increasing involvement. Dynamic Systems in Social Psychology. pp.219-249. Academic Press. (1994).

7. Mukherjee, P., Wong, J. S., Jansen, B.: Patterns of social media conversations using second screens. in ASE BIGDATA/SOCIALCOM/CYBERSECURITY Confence. pp. 1-5. ASE. Stanford, USA. (2014).
8. Mukherjee, P., Jansen, B. J.: Social TV and the social soundtrack: significance of second screen interaction during television viewing. in Proceedings on International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP-BRiMs). 8393, pp. 317-324. LNCS, Springer. (2014).
9. Latané, B., Bourgeois, M. J.: Experimental evidence for dynamic social impact: The emergence of subcultures in electronic groups. Journal of Communication, **46**(4), pp. 35-47. (1996).
10. Bourgeois, M. J., Bowen, A.: Self-organization of alcohol-related attitudes and beliefs in a campus housing complex: An initial investigation. Health Psychology, **20**(6), pp. 434-437. (2001).
11. Bowen, A., Bourgeois, M.: The contribution of pluralistic ignorance, dynamic social impact, and contact theories to attitudes toward lesbian, gay, and bisexual college students. Journal of American College Health. **50**(2), pp. 91-96. (2001).
12. Harton, H., Binder, D., Russell, E.: Clustering and consolidation in real time computer discussions. in Unpublished manuscript, University of Northern Iowa, Cedar Falls. (2001).
13. Latané, B., Bourgeois, M. J.: Dynamic social impact and the consolidation, clustering, correlation, and continuing diversity of culture. Blackwell Handbook of Social Psychology: Group Processes, Editors, Hogg, M.A. and Tindale, R.S. pp. 235-258. Blackwell Publishers. UK. (2001).
14. Richter, L., Kruglanski, A. W.: Motivated closed mindedness and the emergence of culture. The Psychological Foundations of Culture, Editors, Schaller, M. and Crandall, C.S. pp. 110-131. Psychology Press. Mahwah, New Jersey. (2003).
15. Lavine, H., Latané, B.: A cognitive-social theory of public opinion: Dynamic social impact and cognitive structure. Journal of Communication. **46**(4). (1996).
16. Henrique, J.: Get old tweets-python computer API. Available at https://github.com/Jefferson-Henrique/GetOldTweets-python. (2017).
17. Ward, G.: Moby Word Lists. Gutenberg Literary Archive Foundation. Oxford, Missouri. (2002).
18. Tsvetovat, M., Kouznetsov, A.: Social Network Analysis for Startups: Finding connections on the social web. 2011 " O'Reilly Media, Inc.". (2011).
19. Blondel, V. D., et al.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory Experiment, **2008**(10), pp. 1-12. (2008).
20. Arlitsch, K. J..: Semantic web identity of academic libraries. Journal of Library Administration. **57**(3), pp. 346-358. Taylor & Francis. (2017).
21. Jacomy, M., et al.: ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One. **9**(6). pp. 1-12. (2014).
22. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. in Proceedings on Third International AAAI Conference on Weblogs and Social Media. pp. 361-356. AAAI. (2009).
23. Arnaboldi, V., et al.: Online social networks and information diffusion: The role of ego networks. Online Social Networks Media. **1**, pp. 44-55. Elsevier. (2017).