Community Structure Detection and Analysis in Disaster Related Tweets

Harriet Angelie A. Gonzales
Department of Computer Science
University of the Philippines Cebu
Gorordo Ave, Lahug, Cebu City
harrietangelie@gmail.com

Kurt Junshean P. Espinosa
Department of Computer Science
University of the Philippines Cebu
Gorordo Ave, Lahug, Cebu City
kpespinosa@up.edu.ph

ABSTRACT
One of the considered principal disasters that hit the Philippines almost year round is flooding. At the occurrence of such floods, social media – Twitter for instance – serve as communication outlet between users rendering them significant in information gathering and dissemination. This study aims to determine the significance of social networks when it comes to disaster information by analyzing community structures formed from different graph relationships and comparing it to actual patterns of flood affected areas of the same timeframe. This paper analyzes the properties of the community structure detected among nodes in a social network graph formed among Filipino Twitter users who tweeted about flood. Interaction relationship graph was created wherein an edge is formed between two users if user A mentions user B. Seventy-seven communities with more than ten nodes were detected. However, nodes belonging in the same community did not show similarities with each other.

Categories and Subject Descriptors
E.1 [**Data Structures**]: Graphs and Networks.

General Terms
Algorithms, Languages.

Keywords
Graphs, Community Structure, Social Networks, Data Mining, Graph Analysis.

1. INTRODUCTION
Social media has become increasingly influential as years passed by. In a recent survey[1] last May 2013, statistics showed that 72% of the adult population is a social networking site user. Because of such influence and population reach of online social networks (OSNs), researches that are in line with social network analysis are becoming more and more popular. One of the recently becoming an area of interest is community detection in online social networks. Community detection is about finding similarities between highly dense network relationships. It is believed that a network area more concentrated than the rest is said to be forming a community. And like the real world, people belonging to the similar community will exhibit similarities in behavior which highly interests social researchers recently. In the past few years, government institutions in the Philippines finally realized the power of social media that a lot of the government institutions have corresponding twitter accounts where they post announcements. Noticeable among these announcements are those that are disaster-related. Giving out warning signals, precautionary measures, and even detailed information such as amount of precipitation, social networks are becoming a medium for community awareness. Because of this, to know the significance and impact of social networks in disaster information is really an interesting topic. To be able to analyze noticeable patterns between network nodes will really help in improving approaches in information dissemination in social networks.

2. RELATED WORK
According to a primer[2], the Philippines is considered as a disaster hotspot. Tropical storms and floods, with 102 and 72 occurrences respectively for year range 2000-2012, are considered to be principal disasters[3]. Flooding is one of the frequently occurring natural disasters in the country and the damage is highly significant. Because of this, a lot of effort has been put into researches that aim to mitigate the damage of frequent flooding in the country.

A lot of studies use social networks to discover certain behaviors of people. A particular study by Lee et al. [2] focused on the different behavior of tweeter users when tweeting under the circumstances of an occurring disaster, specifically flooding. They did a classification on "participant" and "observer" tweets in order to find common behavior among "participants" (i.e., those twitter users who are within the vicinity of flooded areas). This study is quite similar to theirs in terms of processing disaster related tweets and analyzing the behavior of Twitter. However, in terms of behavior analysis, this paper is geared towards creating graphs and detecting community structures to find similarities in behavior among tweets which belong in one community, and

[1] The demographics of social media users: http://www.pewinternet.org/2013/02/14/the-demographics-of-social-media-users-2012/

[2] World Bank's Primer on Reducing Vulnerabilities to Disasters: http://siteresources.worldbank.org/INTEAPREGTOPURBDEV/Resources/Primer_e_book.pdf

[3] Data from the Senate Economic Planning Office: http://www.senate.gov.ph/publications/AAG%202013-04%20-%20Natural%20Disasters_final.pdf

Philippine Computing Journal Dedicated Issue on Natural Language Processing, pages 11–16
Vol. IX No. 1 August 2014

making use of these structures to find a way of solving Twitter's scarcity of geolocated data.

To analyze data gathered from Twitter, patterns must be found within the network of tweeters (users). Social network analysts use graphs and matrices to represent information about pattern ties among social networks [1]. Nettleton [3] said that social network analysis has recently experienced a surge of interest due to different factors such as popularity of online social networks (OSNs), their representation and analysis as graphs, the availability of large volume of OSN log data, and commercial/marketing interests. According to him, graph mining in online social networks, though relatively new to research area, is firmly based on the basic concepts of graph theory and relational concepts of individual interactions, how they interrelate, group together and follow each other.

Finding communities in networks has recently been of considerable interest [4]. Communities, as he described it, are groups of vertices within which connections are dense, but between which connections are sparser. In real word, communities are the groups we form such as our friends or family or even the working environment. Community detection in social networks is a line of analysis on networks to provide further conclusions [2].

There are different existing algorithms for community detection among networks. A lot of studies have been made to compare algorithms between each other. In the study of [2], they also touched on community detection in which they used the Walktrap algorithm. This is a new algorithm that captures much information compared to previously proposed algorithms for community detection.

The efficiency of Walktrap algorithm is well presented in a journal article by Pons and Latapy [5]. In their paper they compared quality and time efficiency of eight community detection algorithms specifically, Fast Modularity, Donetti Muñoz, Cosmoweb, Girvan and Newman, Netwalk, Duch Arenas, MCL, and of course Walktrap.

3. METHODOLOGY

This study aims to find relationships between nodes in social networks to assess the significance of social networks in mapping disaster information. Because social networks such as Twitter are widely used today, this study wants to find entry points as to where social network updates could prove to be holding significant information in the context of disasters, floods specifically. To prove such significance, data from social network updates must be able to show which places are flooded. Being able to tell such information will indeed prove that social networks as of today could be significant data sources. However, problem with most tweets is that it lacks detailed information. Because of this, instead of getting information from the actual tweets themselves, the researchers studied and tried to find relationships based on the similarities between tweets and looked for certain patterns that would suggest a flooded area and because of this, graphs were utilized.

This study represented its dataset as weighted graphs with edge weights depending on the relationship type, giving nodes with higher probability of being flood victims, a greater edge weights. A particular property of graphs that is observed and analyzed in this study is the formation of communities among networks. In the scope of this study, an edge is formed between users when user *A* mentions user *B*. Since graph used is unidirectional, then a symmetrical matrix is formed wherein when there is an edge $A \rightarrow B$, there must also be an edge $B \rightarrow A$.

3.1 Data Gathering

Twitter streaming was implemented using Twitter4J[4] library. A request filter was implemented with the following parameters used:

Track – {flood, flooding, flooded, baha, nagbaha, bumabaha, bumaha, binaha}

Location – {4.468110, 116.812721, 21.234150, 126.856628}

However, no tweets on flood were gathered since it is already past rainy season. To be able to proceed with the experiments, a different dataset [2] was used. It consists of 600,000 tweets gathered from August 6 – 9 during the flooding brought about by Habagat last 2012.

3.2 Data Filtering

Since the dataset did not come through the streaming API and through the request filter, a pre-filtering has to be done. Tweets which did not contain any track keyword were filtered out which then left 162,088 relevant tweets for this study. Sample data were then selected from the remaining tweets. In selecting sample users, it was checked whether such account still existed since it was a relatively old dataset. Since the Habagat dataset consisted only of database id, tweet id, tweet string, username, timestamp, and retweet count, additional queries were done using the Twitter4J library. Each sample user was classified either as a participant or as an observer. Participant users are those with firsthand experience such as:

- baha sobra! Above waist level na ata... 'di ko na macheck lakas ng ulan eh. Sinara na namin lahat ng doors and windows (English Translation: *Heavy flooding! It seems it's already above waist level… I can no longer check because it's raining hard. We already closed all the doors and windows*)
- The flood's inside na!!!!!!!!

If a tweet containing the track keywords is a retweet or is not a firsthand experience then it is labeled as an observer tweet:

- RT @mitchsanchez: PLS RT: Water rising to 2nd flr. 22 b evangelista st xavierville 1 qc. Neil Flores. 20 people in the house. Please sen ...
- Pray for the people who were and still affected by the flood. :(#PrayForThePhilippines

3.3 Social Network Graph Building

As proposed, an interaction relationship graph is generated. There is a coefficient x for any node such that x = 0.8 if node is a source and x = 0.4 if otherwise leading to a node relationship weight $nr(i,j) = x_i\ x_j$ for any two nodes i and j. Assigned coefficient values are hypothetical and may be varied to find optimal coefficient. Assignment of coefficient values independent of the graph relationship serves the purpose of

[4] Twitter4J – A Java library for the Twitter API: http://twitter4j.org/

giving more value on source – source relationships which, in would increase the probability of that particular community to be flood victims.

In an interaction relationship graph, nodes are graphed according to retweets and mentions occurring between two nodes. Users tend to mention in their tweets only those whom they are acquainted with, with the exception of mentioning public personalities. Because of this, the idea of finding patterns of interactions between any two nodes retrieved from the dataset might suggest a neighboring relationship parallel to the real world.

An n x n adjacency matrix T is produced where n is the total number of nodes (users). T_{ij} is the interaction weight for any two given nodes i and j. Nodes i and j are neighbors if and only if i mentions or retweets j, or vice versa. An edge $e(i,j) = 0.5$ for retweets and $e(i,j) = 0.7$ for mentions such that $T_{ij} = e(i,j)$ x $nr(i,j)$.

3.4 Graph Creation, Community Detection and Visualization

The application is done in R Language. This is to accommodate convenience in community detection and visualization. The igraph library[5] for R is used which incorporates graph creation, community detection, and graph visualization. The *Walktrap* algorithm in igraph is used for community detection.

4. RESULTS

Three experiments were done with first two experiments showed to be a failure and unsuccessful in detecting communities.

For Experiment 1,400 tweets were randomly chosen from the file of filtered tweets. Sample size was obtained using the formula for known population size. Using Java randomizer, 400 unique numbers ranging from 1 to 162,088 were collected. The numbers then correspond to the line number of the tweet to be included in the sample data. Out of the 400, only 153 users were left after those user accounts that can no longer be traced were discarded. Table 1 and Figure 1 show the data count and interaction relationship graph for Experiment 1.

Table 1. List of corresponding data count for Experiment 1

USERS	153
TWEETS	153
EDGES	52

[5] Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

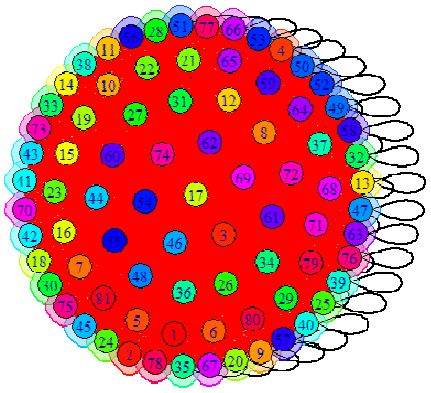

Figure 1. Community Detection in Interaction Relationship Graph consisting of 153 nodes and 52 edges

The low number of sample data might have caused the inability to detect communities and therefore the data is increased to 1000 nodes for Experiment 2. This time, data wasn't chosen in random. The filtered dataset file is arranged according to time of tweet post. Randomly choosing the data in the first experiment might have lead to totally unrelated tweets. Therefore, the tweets were sampled in this experiment in order. To make sure that there will exactly be 1,000 usable tweets, a tweet will only be included in the sample space if the user account is confirmed to be active.

Out of the 1,000 unique user tweets, 278 were participant and the rest are observers. For the mention relationship on the other hand, 155 edges were detected, 12 of which are participant-participant mention, 44 participant-observer (vice versa), and 99 observer-observer tweets (See Table 2 for summary of data count). When the adjacency matrix was produced, interaction graph created, and community detected, the same happened as with experiment 1, there wasn't any community detected.

Table 2. List of corresponding data count for Experiment 2

USERS	**1,000**	**306** Participant
		694 Observer
TWEETS	**1,000**	**18** Geolocated
		1103 Geolocation Disabled
EDGES	**1,121**	**99** Observer-Observer
		44 Observer-Participant
		12 Participant-Participant

For this experiment, the data was increased for the last time which already included all the tweets in the dataset. Same processes in Experiment 2 were followed except for the tweet classification (See Figure 2 for the flowchart). For the first two experiments, tweet classification was manually done however, for this experiment, since data has become too large it is very impractical to do manual tweet classification. Therefore, an automatic tweet classifier is implemented. The tweet classifier is based on some loose lexical rules implemented by the researcher. Using

LangDetect[6] library for Java, tweets are determined whether it is written in English or not because different lexical rules apply respectively (See Figure 3 and Listing 1).

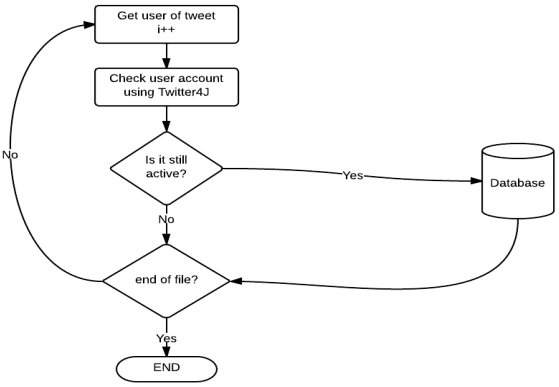

Figure 2. Experiment 3 Flowchart

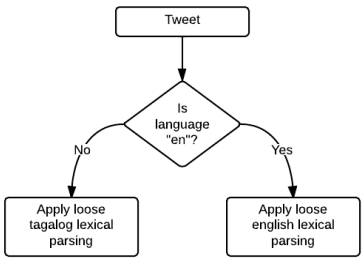

Figure 3. Language Detection for Automatic Tweet Classification

For Tagalog tweets, the string must be cut into three parts (tokens) where the left token is a substring starting from characters at index 0 until the character at the left of a track keyword present in the string while the right token is a substring starting from the character on immediate right of that of a track keyword (See Figure 4).

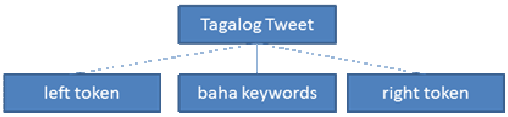

Figure 4. Keywords/Indicators for Lexical Parsing

Each token then in the Tagalog tweet has different lexical rules to comply in order for the whole string to be classified as either a participant or an observer tweet. For English tweets, however, keywords or indicators are subjected to the whole string.

TAGALOG KEYWORDS
negators
["no", "wala", "walang", "wlang", "wlng", "walng", "la", "hindi", "hinde", "ndi", "inde", "indi", "di", "hnd", "nawala", "humupa"]
uncertainty
[daw","raw","parang"]
indicator
[umabot","abutan","aabutan", "papasukin", "pinasok", "pumasok"]

ENGLISH KEYWORDS
reliefEffort
["victim","donate","donation","donating", "volunteer", "relief effort", "relief drive", "rescue", "help"]
relayedTweet
["told","said","report","according"]
negativeKWEng
["not","no","might","seems","maybe","possibly","subsided","if", "philippine", "update", "updates"]
positiveKWEng
["i","i'm","our","us","house","home"]
floodPhrases
["flood here","flooding in"," is flooded","flooding at","by flood"]
questionKW
["who","what","when","where","how","why"]

Listing 1. Keywords/Indicators for Lexical Parsing

After classifying all the tweets, edges were determined for the Interaction Relationship graph. For this experiment, there is an edge between two users if user *A* mentions user *B*. Since Interaction Relationship graph is symmetrical, if an edge is formed from user *A* *to* user *B* when user *A* mentions user *B*, an edge is also automatically created for user *B* to user *A*.

Table 3. List of corresponding data count for Experiment 3

USERS	**35,486**	**15,654** Participant
		35,740 Observer
TWEETS	**51,395**	**690** Geolocated
		50,075 Geolocation Disabled
EDGES	**49,181**	**36,769** Observer-Observer
		0 Observer-Participant
		12,412 Participant-Participant

However, when the adjacency matrix was created, it invoked an Out of Memory error and cannot finish the process because of too large dataset. Therefore, instead of creating an adjacency matrix, list of edges were automatically fed into the igraph library. Out of 49,181 edges, 9,794 communities were found, 77 of which consisted of more than ten nodes (See Table 4; note that numbers in bold style are the community numbers).

[6] LangDetect. Language detection library for Java. https://code.google.com/p/language-detection/

Table 4. Communities and Corresponding Sizes (number of users belonging in one community)

1	41	27	10	115	10	515	11
2	23	28	11	131	10	540	10
3	26	30	15	139	10	616	27
4	14	32	14	144	12	617	24
5	13	33	14	154	10	618	15
6	11	34	12	156	10	619	12
7	11	35	11	158	27	620	11
8	11	36	12	169	18	621	11
9	10	42	10	172	10	622	10
10	13	43	10	174	10	623	10
11	11	45	10	175	10	624	10
12	17	59	11	184	10	629	11
13	11	61	12	211	11	631	10
14	11	64	17	244	10	677	11
16	10	66	24	278	10	736	10
17	14	68	13	305	11	9769	20168
18	12	74	10	321	10	9787	11
19	10	81	18	322	10		
21	13	98	13	412	12		
24	10	111	11	468	10		

5. DISCUSSION

It was found that among the tweets, data on participant – participant mentions were found. In real life networks, people tend to only concern themselves, in times of disasters, on those people whom they have certain connection (e.g. neighbor, friends). Since social networks are said to imitate a real world network, a person who specifically mentions other person in their tweets about floods, on average cases, can be said to be at least an acquaintance of that mentioned person. Finding communities – *a mentions b, b mentions c, c mentions a, d mentions b, a mentions d* for example will most likely yield to *a,b,c,d* getting classified into one community in an interaction-relationship generated graph. If *a,b,c,d* happens to be users classified as participants (flood victims) and they belong in a single online social network community, there will be a high possibility that they also belong in a single community in real life (by virtue of the concept of social networks). Since most real life communities are brought about by closeness in vicinity, this will aid in locating occurrences of floods. This is because a number of people talking about floods coming from the same community will most likely suggest that the community from which they belong has been affected by floods. This concept is a workaround for twitter's scarcity of data in terms of geolocation. This works in such a way that if *a* is the only one who enabled the geolocation feature in Twitter, it can be verified and concluded that *a's* location has been subjected to flooding when *a* will belong in a community in which the majority of nodes are classified as participant nodes.

As for the experiments done, inability to detect communities for experiments 1 and 2 could be attributed to the scarcity of data. Because the data is too scarce, the edges created among nodes were also scarce resulting to a sparse adjacency matrix. Because experiment 3 has been created out of edge lists, detection of communities became possible because the unnecessary data, as that of creating adjacency matrices, has been eliminated. Out of all the communities, four random numbers was generated to get sample communities from which similarities among users will be analyzed. Communities 3, 32, 144, and 617 have been chosen. Based on the results (See Table 5, Table 6, Table 7, and Table 8 for samples), out of 26 users that belong to community 3, only four of them were participant users and the rest are either observer or a mentioned user. Mentioned users that do not belong to the original list of users from the dataset for experiment 3 were removed since no data were gathered from them. There is really no visible similarity in the context of the tweets among the users in the same community for communities 3, 32, and 144. However, it was noticed that most of the mentioned users in the tweets are found to belong within the same community as those users who mentioned them. Community 617 however showed a similarity in the content of the tweets themselves. In community 617, it is composed of users who retweeted the same tweet.

Looking at the profile location of the users in the same community, it is really varied. It could not really be inferred that they come from a place of close proximity. This is however subject to the unreliability of data that the users type in their profile location.

Table 5. Sample users that belong to Community 3

Users	Tweets	Classification
User 3963	Todo bigay yung baha. Nararamdaman kong wala nang one month sembreak! :(	P
User 3965	@edeeeeeeeeen oo nga :| eh hindi kasi ako makapagfocus eh. Bothered by the flood situation dito sa bahay tapos buong araw nakatutok sa news	O
User 17457	@nicables I have. Mga noodles. But im still hoping na makakaen ng meal. Haha. Arte ko! :)) Nasa dorm ako. Knee deep pa ata baha	P
User 21252	@angeliqueortiz Grabe baha sa street ng dorm :O	P
User 31267	@vionnana27 @nicables nung tuesday kaya lumusong ka ng baha. :))	P
User 44999	RT @angeliqueortiz: Sana mawala na yung flood sa Espana :(gusto ko na umuwi :((	O

Table 6. Sample users that belong to Community 32

Users	Tweets	Classification
User 897	@andyr0704 hindi ko pa na try na pasulungin sila sa baha. Hindi sila sanay and I can feel na stressed na sila. Plus ginaw and gutom. :(	P
User 899	Becky housemate nag pasama sa katapat na tindahan kahit hanggang singit at ang baha kasi may gwapong nag yoyosi!	P
User 2038	@jopaydiets baha pa din makati?	O
User 2039	@paulcahanding Na baha sa inyo!!!! I forgot! Navotas ka dba??? :(	O
User 12307	@yumitamazaki seryoso??? Hahaha! Ano buzz... Eh mahirap talaga network dahil sa ulan at baha. Laughtrip yan ah!!!	P
User 14734	@paulcahanding @justviewing13 busy ata baha kasi at ala kuryente hehe	P

Table 7. Sample users that belong to Community 144

Users	Tweets	Classification
User 10608	@shannyLae Sana wala pa. kasi ung north quadrangle baha pa.	P
User 28281	@monnnmonnn @tristansy @_ivannnnn @OlginaRenz @kowalsskkii @awwwkDWARD @SuperRenceeee haha.. flood ba YIKES . sensya	O
User 28283	RT @YouCanKissMe: @tristansy @awwwkDWARD @_ivannnnn @kowalsskkii @OlginaRenz bat ako naka tag tae pag bukas ko FLOOD baha sa wall ko h ...	O
User 28284	Banat naman kayo sa'kin para mawala lungkot ko dahil sa baha :'(	P
User 42519	RT @_ivannnnn: @awwwkDWARD @AkikoGuevarra1 Hahaha buti nga wala ng baha dito sa'min huminto na rin kasi yung ulan :))	O
User 44511	@_ivannnnn pinasok po kau ng baha?	P

Table 8. Sample users that belong to Community 617

Users	Tweets	Classification
User 33	RT @LeeCuriosity: Remote Sensing Mast deployed successfully! Starting to get a flood of new Gale Crater imagery from #MSL #Curiosity! ht ...	O
User 65	RT @LeeCuriosity: Remote Sensing Mast deployed successfully! Starting to get a flood of new Gale Crater imagery from #MSL #Curiosity! ht ...	O
User 507	RT @LeeCuriosity: Remote Sensing Mast deployed successfully! Starting to get a flood of new Gale Crater imagery from #MSL #Curiosity! ht ...	O

6. CONCLUSION

As mentioned by Pons and Latapy [5], *Walktrap* algorithm indeed showed the best performance in terms of correctly determining the two communities out of the dataset. Other community detection algorithms incorporated in the igraph library failed to detect correctly the expected number of communities. However, community detection as an algorithm for clustering node similarities pose to be highly dependent on the type of data it is subjected. Based on the experiments created so far, it seems that despite it being a good clustering algorithm for large real life networks, it will be very difficult to detect communities among nodes when data are too scarce for the kind of relationship graph being created. To efficiently detect communities, graph creation from edge lists is recommended because creation of adjacency matrices might lead to sparse matrices which, in a way, render community detection difficult. Also, finding actual flood locations using an Interaction Relationship graph proved to be a not probable solution. However, community detection is indeed efficient in clustering data but the type of data clusters that is aimed to be achieved is dependent of the type of graphs created – that is to say, the criteria for two nodes to form an edge must be well selected.

7. RECOMMENDAITON

The concept of using community detection to identify the different affected locations proved to be of significance. However, the results gathered were incorrect because the basis for similarity between graph nodes showed to be an incorrect approach to achieve the desired results. To improve this study, the researcher would suggest that natural language processing methods must be used. That is to say that the nodes must be tweets themselves and not the users of the tweets. It will then follow that the basis for the similarities between nodes must be based on the context of the tweet themselves. Furthermore, the researcher suggests that a better relationship graph and weighing criteria must be looked into. NER (Name-Entity-Recognition) field in Natural Language Processing is a probable solution in creating better similarity graphs by extracting locations themselves from tweets and creating similarity matrix based on similarities in location detected. Categorizing between participant tweets and observer tweets can be improved by using NLP methods to determine whether the tweet talks about a first-hand flood experience or not. Communities formed after improving the methods as described above stand a high possibility that the locations produced where participant tweets comprise the majority translate to having these locations be actual areas of disaster in real life.

8. REFERENCES

[1] Hanneman, R., and Riddle, M. 2005. Introduction to Social Network Methods. Riverside, CA: University of California.

[2] Lee, J. B., Ybañez, M., De Leon, M., and Estuar, M.R. 2013. Understanding the Behavior of Filipino Twitter Users during Disaster. International Journal on Computing, 3(2).

[3] Nettleton, D. D. 2013. Data Mining of Social Networks Represented as Graphs. Computer Science Review, 7: pp. 1–34.

[4] Newman, M. E. J. 2004. Detecting Community Structure in Networks. The European Physical Journal B - Condensed Matter and Complex Systems, 38(2): pp. 321-330.

[5] Pons, P. and Latapy, M. (2006, April). Computing Communities in Large Networks Using Random Walks. Journal of Graph Algorithms and Applications, 10(2), 191-218.