Real Time Social Network Data Analysis for Community Detection

Mohammad Sarwar Jahan Morshed
Department of Computer Science
American International University - Bangladesh

Akinul Islam Jony
Department of Computer Science
American International University - Bangladesh

ABSTRACT
In WWW becomes a widely used platform for different social networks and social medias for the social communication. This platform becomes the oasis of a huge amount of data. Therefore, this data repository draws tremendous attention from corporate, government, NGOs, social workers, politician, etc. to either promote their products or to convey their message to the targeted community. But identification of community structure and social graph becomes a challenging issue for the social network researcher and graph theory researchers since the pervasive usage of instant messaging systems and fundamental shift in publishing contents in these social medias. Although a lot of attention has been given by the researcher to introduce several algorithms for identifying the community structure, most of them are not suitable for dealing with the large scale social network data in real time. This paper presents a model for community detection from social graph using the real time data analytic. In this paper, we introduce data analytic algorithms that can analysis contextual data. These algorithms can analyze large scale social interaction data and can detect a community based on the user supplied threshold value for community detection. Experiment result shows that the proposed algorithms can identify expected number meaningful communities from the social graph.

Keywords
Real time data, Social Network, Community Detection, Big data, Data Analytic.

1. INTRODUCTION
We Social networks have made a major transform in recent days. These networks form a decent approach to share instantaneously a variety of information between individuals and their neighbors in the social graph. Online social networking sites present an actual image of the structure and state of affairs of the society as well as the interaction of the generation with technologies and others. These days users of social networks both create and consume a significant extent of social content. More than 200 social networking sites performing worldwide and this number is growing [1]. This shows that social networking has formed a new age where content sharing through Social networking sites is an everyday practice. People use social networks not only from their laptops or personal computers, but also from their handheld devices. Therefore, these sites become useful tools for staying in touch with friends, family, neighbors, and colleagues for sharing things instantly. Random use of the social network sites creates the ocean of data.

On the other hand, real time data analytic has huge impact on research, business and society. In research, data analytic turns information into insight, connects and empowers people and community, makes facts based assessment, and builds integrated solution to multifaceted problems. In IBM UIDP conference it is revealed that businesses that apply data analytics are four times more likely to outperform others. Also data analytic can play a vital role for agricultural development, weather forecasting, secure deep earth mining, critical infrastructure/facility protection, environmental monitoring, analysis web data or e-commerce, telecom data analysis, purchase data analysis at department/grocery stores, bank/credit card transactions analysis, national security, etc.

Then why and when community detection is required for community detection? Followings can be some of the usage of real time social network data analysis:

- Individual can analyze real time data for his personal purposes.
- To provide social security by finding terrorist instantly by analyzing their social and communication data in social network, so that bloody terrorist attack can be prevented.
- Finding appropriate person (for instance specialist in different occupation) in order to engage him during national crisis, such as earth quack, natural disaster, famine, terrorism, etc.
- For commercial purposes to send offer of a particular product to maintain the loyalty of the existing customer by providing different kinds of offer.

Although large volumes of data are stored in the Social Networks' Server, there is no value of them unless used properly. In this paper we present a mechanism that correlates three important issues such as (a) social networks data, (b) real time data analysis, and (c) community detection. Therefore, the proposed approach could be an effective tool for solving social, commercial, individual, medical problems by using social network data.

Outline of this paper is: in section 2, related works have been stated. Section 3 presents the proposed model while section 4 illustrates the implementation of the working procedure of the proposed architecture. Simulated results and discussion have been presented in section 5. Finally, conclusion and future works have been stated in section 6.

2. RELATED WORKS
All Social Network is one of the most promising recent trends in the Internet. Data collection from social network can be done in different ways where one of the easiest ways is crawling tools. Besides, data collection can also be done by using special kind of API (Application Programming Interface). But it is necessary to consider the category of data while performing data collection. Usually there are three categories of digitally-encoded data which includes public (can be accessed by everyone), semipublic (can be accessed by small group of individuals), or private (can be accessed only the owner) [2]. In the case of semipublic and private,

authentication and authorization is needed for accessing data. Public data can easily be collected by the search engine or by the crawler. But hidden data (semipublic and private) cannot be collected easily by the crawler.

Therefore a lot of research attention is paid for monitoring/collecting/analyzing the data from the social Network sites. Doan et al. [3] presented an approach for analyzing Twitter messages and they illustrated their approach by tracking influenza-like illnesses as an example. Similarly Earle et al. [4] also analyzed Twitter messages but for monitoring earthquakes. Several data analysis research works have been presented for studying data in the era of political environment, such as political micro blogging data analysis in Sweden by Larsson and Moe [5] and data (tweets from twitter) analysis during elections in Singapore by Poor et al. [6]. Aliprandi and Marchetti [7] presented a collaborative platform CAPER for the purpose of preventing organized crime where CAPER is capable of Open and Closed Information Acquisition, Processing and Linking.

There are number of research which is devoted to data crawler tools. A crawler is a sort of software or tools which is capable for crawls over web data. In several existing article crawler is also named as "web robots" [8] and "web spider" [9]. Crawlers are widely used in popular search engine such as Google or Yahoo. Crawling process starts by taking a list of URLs and fetching information from that web pages and next crawlers extracts URLs from the fetched web pages if they found any URL, then again fetched information from that web pages until all the list of URLs' fetching is complete. Besides, for collecting data a crawler needs to update data periodically. In 2004 Yih et al. [10] presented an incremental crawler for collecting and periodically updating data of a forum. Besides this, there many research works on crawler, such as Bergholz et al. [11] describe a hidden web crawler, Duda el al. [12] describe a crawler that deals with asynchronous JavaScript and XML (AJAX), Peng and Wen-Da [13] describes a crawler to crawl information in a financial field, and Yang and Hsu [14] describes focused crawling of sites with calls for papers. But in our paper we have used API-based approach for collecting data from the social network.

Social networks contain a community structure in which communities are groups of nodes in the network, tightly connected with each other [15]. There are a number of methods for community detection in social network, such as the Kernighan-Lin algorithm [16], and divisive algorithm [17].

Papadopoulos el al. [18] surveyed the performance characteristics of community detection algorithms. But in this paper, not only the static data but also the real time data is considered for detecting community in social network which makes much more sense in this regard.

3. PROPOSED MODEL

This paper presents a novel approach for community detection from the social networks' data by real time data analysis. This model comprises five layers as shown in the Fig. 1. Different kind of social networks such as Facebook, LinkedIn, Researchgate, Twitter, Google Plus etc. lie in the bottom layer (*social networks*) of this model.

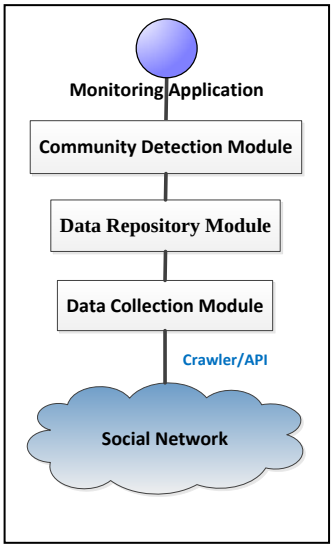

Fig 1: Five-layered Model

Data collection module is placed on the immediate top of this bottom layer (*social networks*). Data collection layer is actually a data mining crawler or API for collecting real time data. Crawler design and development is out of the scope of this paper. A crawler design is a future work of this proposed model which is already in the implementation phase for the proof of concept. Instead we develop a *data collection engine* based on the native APIs provided by the social networks. This data collection engine establishes connection between Data Connection Module of the proposed architecture and different social networks' provided APIs.

This module collects data from social network provided data using this data engine. User just provides their respective social network credential to access those social networks. API will ask social network users which data will be collected from user's social network. Social Network provided API is used for interfacing relevant social network repository; for instance Graph API is used for collecting individual's Facebook data.

Third layer of this model is *data repository module*. Functionality of this module is to store collected social network data in a unified way. This module generates a data model with the common values considering as the keys and store them in NoSQL (JSON) format. Fig. 2 shows a structure of the data format stored by the Repository Module.

Initially, collected data from each social network stored separately in an individual file with the profile name of the corresponding social network user. A filed is created named as Context Key while processing and storing real time data. Context key holds N number of comma separated Meta information from social networks user data. This Meta data can be user's school name, university name, birth data, email, home town, birth place, father name, mother name, interest, likes, hobbies, context based shared events, or documents, distinguishing information extracted from chat box etc. These Meta data is used to analysis and to monitor user's motivation, trend, behavior, etc.

Fourth layer is *community detection module* that ultimately provides to identify contextual community from the social network. An algorithm have been proposed and implemented for analyzing the context key for monitoring user.

```
"userid": "1462662698",
     "userprofilename": "Soikot Ahmed",
     "username": "soikot.ahmed",

     ..................................

     ..................................

     "friends": [
          {
               "Friendsname": "Matthias BrÃ¶cheler",
               "userid": 210100146,
               "username": null,
               "birthday_date": null,
               "email": "210100146@facebook.com",
               "profileurl":
"http://www.facebook.com/profile.php?id=210100146",
               "education": [],
               "movies": "",
               "current_location": null,
               "interests": "",
               "picture":
"http://profile.ak.fbcdn.net/hprofile-ak-
snc4/27443_210100146_9942_t.jpg",
```


Fig 2: Structure of data format

4. PROCEDURES OF THE PROPOSED MODEL

In the proposed model, we have collected the data by the data collection module. Then we categorize the data in two types include *static data* and *moving data*. Static data is used for determining the individual's connection which actually is vertices in the social graph. Various vertex/node attributes are to be used for measuring edges weight. These weights will ultimately detect the community by clustering sub-graphs from a graph. On the other side moving data (such as like, comments of Facebook) measures weight between two nodes which eventually helps to determine the degree of communication between two nodes. A formal definition of notations and algorithms are presented in the next two subsections respectively, which are used for real time data analysis for community detection in social networks

4.1 Notations

In this subsection we have introduces some terms and notations used for real time data analysis. Below is the list of formal definition of some terms and notation in terms of graph and set theories:

- ***G = (V, E)*:** A social network graph where ***V*** represent a set of *n* node/entity such as event, document, individual, etc. and ***E*** represent a set of m edges/connections between two nodes or entities.
- G_i Any individual's social graph where $G_i \in G$.
- ***I***: Set of individuals.
- I_i: Any individual in social network where $I_i \in I$.
- ***C***: Set of connections for individual I_i where $C_i \in C$.
- K_i: Set of context-keys of individual I_i.
- K_s: Set of keys in any individual's social network graph.
- K_r: Set of reference keys for filtering any individual's context-keys.
- W_r: Reference weight between two nodes.
- W_{ij}: Generated weight of edge between individuals ***i*** and ***j***.
- G_a: Clustered community after associating connection.

4.2 Data Collection and data formation

As we mentioned that the proposed model uses a basic crawler using the social networks native APIs (e.g. Facebook Graph API) that is able to collect data from social networks. This collected data is stored as a NoSQL JSON format in a text file (as shown in Fig 2) and the file is named with the profile name of the user. An additional field named "Context-Key" is also being created while data is collected. The Algorithm 1 is used to create context key. In this respect, a reference key set Kr is used to identify the context keys.

Algorithm 1: Context-keys Generations

Input: G, K_r
Output: K_i

for each G_i **do**
  for each *key* in G_i **do**
    for $x = 0$ to k_r.*length* **do**
      if G_i *[key]* = = k_r*[x]* **then**
        k_i *[index++]* := G_i *[key]*
      end if
    end for
  end for
end for

Algorithm 2: Association of Connections

Input: G, K_i, K_j, W_r
Output: *True/False*

W_{ij} := $(K_i \cap K_j)$
  if $(W_{ij} >= W_r)$**then**
    include G_i and G_i into G_A
  end if

In our implementation, K_r = *{home town, high school, graduate school, occupation, work history, likes, shares, conversation, comments on events, follows, etc.}*. From this algorithm, we find a set of the context keys for any individual in social media or networks. For example, Fig 3 and Fig 4 illustrate the sample of collected data from LinkedIn and Facebook respectively with the context keys.

Algorithm 2 identifies the degree of interactions W_{ij} between two individuals *i* and *j*. W_{ij} is the value generated by intersecting the generated context key of *i* and *j*. This weight is compared with a referenced weight W_r which is actually a threshold value set by the data analyzer. If this threshold value equals to the weight calculated from the interaction between two individuals, this algorithm includes both G_i and G_i into an associate social graph G_A for clustered individuals in the social graph.

LinkedIn Data


```
{
   "username": "Muhammad Sarwar Jaha
Morshed",
   "usereducation":"Rajshahi Cadet
College, Khulna University, KTH Royal
Institute of Technology",
   "userwork": "Research Engineer at
LuleÃ¥ University of Technology",
   "email": "sarwar_rcc@yahoo.com",
   "city": "LuleÃ¥, Sweden",
   "friends": [
         "Friendsname": "Abdullah Al
Hasib",
   "work": "PhD Candidate at NTNU",
   "country": "Bangladesh",
   "city": "Trondheim Area, Norway",
   "picture":
"http://media.linkedin.com/mpr/mprx/
0_8LXTPUy2--kZRBTxhkF-P4gCKv_nUBhxGXrO
PJIy7KBQk1c02iqSjMotYRiW4K3PT66xyYh_KYrq",
publicProfileURL":"http://www.linkedin.com
/pub/abdullah-al-hasib/13/157/651",
"contextkey": "Abdullah Al Hasib,PhD
Candidate at NTNU,Trondheim Area,
Norway,no, Bangladesh, LuleÃ¥, Sweden "

        ..........
        ..........
```


Fig 3: Individual LinkedIn Social Graph

Algorithm 3: Relation value Generation

Input: G_A
Output: *interaction*

interaction := *identify_interaction*($P_1, P_2 \ldots P_z, P_{z-1}, P_{z-2}, \ldots P_{z-n}$)) //where represents different interactions between two individuals

return *interaction*

Algorithm 4: Weight Measurement between Connections

Input: G_A
Output:

for x = 0 to G_A.*length* **do**
  for y = 0 to G_A.*length* **do**
    $weight_{ij}$: = *relation_value_generation* (G_A)
      E_{ij} := $weight_{ij}$
  end for
end for

Algorithm 3 is used to find interaction between two individuals based on the real time streaming data or moving data. ***P*** indicates the moving data. Conversation via chat room, similar data sharing, likes, following, commenting on the events or shared data, etc are considered as moving data. Algorithm 4 calls this function by passing the associate social graph G_A. And finally this algorithm assigns the real time interaction E_{ij} between two individuals.

Facebook Data


```
{
   "username": "Blind Sky",
   "usereducation": " Rajshahi Cadet
College, Khulna University, KTH Royal
    Institute of Technology",
   "userwork": "WSP Group,LuleÃ¥ Tekniska
              Universitet",
    "city": "LuleÃ¥, Sweden",
    "friends": [
    {
     "email":"muzahid.mizan@facebook.com",
     "education": [
     {
        "school": {
           "id": 110799508948308,
           "name": "Noapara Model
              School,Jessore"
           },
         "type": "High School"
     },
     {
         "school": {
            "id": 106053336092138,
            "name": "North South
                    University"
                     ..........
          }
        }
      },
    "contextkey": "Rajshahi Cadet College,
Khulna University, KTH Royal Institute of
Technology, WSP Group,
LuleÃ¥, Sweden"
},
              ..........
```


Fig 4: Individual Facebook Social Graph

5. EXPERIMENTAL RESULT & DISCUSSION

For the proof of concept we have developed a prototype to simulate the algorithms of the proposed model. As we already declare in the earlier section that the implementation of data collection module is out of the scope of this paper. Instead, we have developed a basic crawler using the social networks native APIs (e.g. Facebook Graph API) to collect data. Using this crawler, 5000 of individual's social graph have been collected, formatted in JSON file and stored in our server for data analysis. Experiment shows that number of community detection increases with the increase of data set as shown in the Fig 5. But this increase is not proportional. The maximum number of communities have been detected by our algorithm is 90 from 5000 individual social graph.

This proposed algorithm uses a user supplied reference threshold value for indicating the weight of interaction between the individuals in the universal social graph. This threshold value ranges from 0 to 1. Weight of interaction between two individuals are measured based on their interaction (for examples, conversation, sharing contents, likes, similarities in interest, comments, etc.). If the threshold value increases, number of community detection decreases. In our experiment it shows that the number of detected communities is 58 for a threshold value of 0.5. This result is more than half of the total detected community for the 5000 social data set.

Our initial experiment shows that the algorithms for proposed model produced good result in terms of community detection. Efficiency and effectiveness of a real time data analytic depends on data collection, data analysis, and a prompt response based on the data analysis. So the limitation of this model is that a crawler for real time data collection is yet to be

implemented. Besides, more data sets need to be sampled for proving the scalability of the proposed data analytic algorithms.

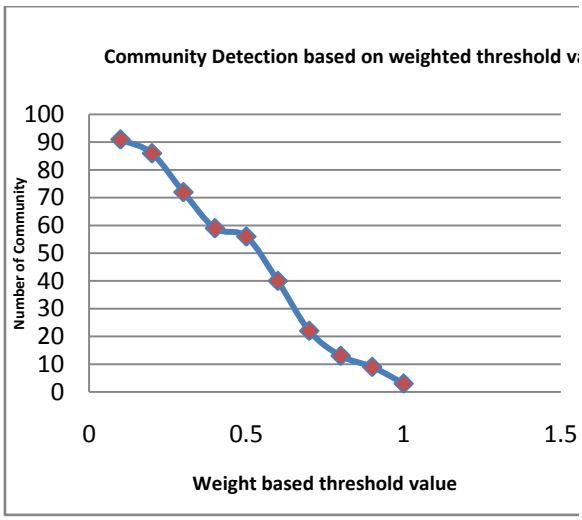

Fig 5: Community detection from the social data graph

6. CONCLUSION

Social networks become a universal data repository where large volume of data is being stored in every moment. This huge volume of data becomes the focal point for corporate, Governments, NGOs, individuals, social workers, politicians, etc. to expose their products or messages to the targeted community. Therefore, identifying relevant community from the social graph of the social media becomes a challenging issue for the researcher. In this paper, we present a model that can detect community by real time data analysis. Algorithms used in this model already show a promising result for community detection. In this paper, we used 5000 social data set from different social networks. But simulation should be performed on more number of data set - which has actually been left open as the future work. Besides, implementation of a complete crawler for real time data collection from the social networks/medias has also been left open as a future work.

7. REFERENCES

[1] Wikipedia, "List of social network site", http://en.wikipedia.org/wiki/List_of_social_networking_websites", 2014.

[2] Semenov, A. and Veijalainen, J. A modeling framework for social media monitoring, in IJWET, in press, 2012.

[3] Doan, S, Ohno-Machado, L., and Collier, N. 2012. Enhancing Twitter Data Analysis with Simple Semantic Filtering: Example in Tracking Influenza-Like Illnesses. In IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB). Ohno-Machado, Lucila, Jiang, X. (Eds.), California, USA: IEEE Computer Society, pp. 62 –71.

[4] S. Earle, P., C. Bowden, D. and Guy, M. 2011. Twitter earthquake detection: earthquake monitoring in a social world. In Annals of Geophysics, vol. 54(6), pp. 708-715.

[5] O. Larsson, A. and Moe, H. 2012. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. In New Media Society vol. 14(5), pp. 729–747.

[6] Skoric, M., Poor, N., Achananuparp, P., P. Lim, E. and Jiang, J. 2012. Tweets and Votes: A Study of the 2011 Singapore General Election. In 45th Hawaii International Conference on System Science (HICSS), 2012, pp. 2583 – 2591.

[7] Aliprandi, C. and Marchetti, A. 2011. Introducing CAPER, a Collaborative Platform for Open and Closed Information Acquisition, Processing and Linking. In Stephanidis, C. (Ed.), HCI International 2011 – Posters' Extended Abstracts. Berlin, Heidelberg: Springer, pp. 481–485.

[8] Heinonen, O., Hätönen, K. and Klemettinen, M. 1996. WWW Robots and Search Engines. In Seminar on Mobile Code No. TKO-C79), Helsinki University of Technology, Department of Computer Science.

[9] Thelwall, M. 2011. A web crawler design for data mining. In Journal of Information Science, vol. 27(5), pp. 319–325.

[10] Yih, W., Chang, P. and Kim, W. 2004. Mining Online Deal Forums for Hot Deals. In Zhong, N., Tirri, H., Yao, Y., Zhou, L., Liu, J., Cercone, N. (Eds.), Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, WI '04. Washington, DC, USA: IEEE Computer Society, pp. 384–390.

[11] Bergholz, A. and Childlovskii, B. 2003. Crawling for domain-specific hidden Web resources. In Santucci, G., Klas, W., Bertolotto, M., Calero, C., Baresi, L. (Eds.), Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. Washington, DC, USA: IEEE Computer Society, pp. 125 – 133.

[12] Duda, C., Frey, G., Kossmann, D., R. Matter, D. and Zhou, C. 2009. AJAX Crawl: Making AJAX Applications Searchable. In Ioannidis, Y.E., Lun Lee, D., Ng, R.T. (Eds.), Proceedings of the 2009 IEEE International Conference on Data Engineering, ICDE '09. Washington, DC, USA: IEEE Computer Society, pp. 78–89.

[13] Peng, L. and Wen-Da, T. 2010. A focused web crawler face stock information of financial field. In Zhou, M. (Ed.), 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), pp. 512 –516.

[14] Y. Yang, S. and L. Hsu, C. 2009. Ontology-supported web crawler for information integration on call for papers. In Chan, P. (Ed.), International Conference on Machine Learning and Cybernetics, pp. 3354 –3360.

[15] Girvan, M. and E. J. Newman, M. 2002. Community structure in social and biological networks. In PNAS, vol. 99(12), pp. 7821–7826.

[16] Kernighan, B. and Lin, S. 1970. An Efficient Heuristic Procedure for Partitioning Graphs. In The {B}ell system technical journal, vol. 49(1), pp. 291–307.

[17] E. J. Newman, M. 2004. Fast algorithm for detecting community structure in networks. In Phys. Review, vol. 69(6).

[18] Papadopoulos, S., Kompatsiaris, Y., Vakali, A. and Spyridonos, P. 2011. Community detection in Social Media. In Data Mining and Knowledge Discovery, vol. 24(3), pp. 515–554.