Scraping Global Threats in Facebook Through Movement Patterns Generated by Random Walks

Pratishtha Mishra, Prachi Jain, NabanitaMajumder and
Shivani Paul
Department of Computer Engineering
University of Pune

Abstract *– The increased concern of national security agencies due to the occurrences of emerging global threats through open social networks (OSN) has been significant. The risks have occurred in large measure due to explosive growth of social media, which facilitates immediate worldwide engagement. We have studied a set of four schemes based on random walk exploration to crawl across the non-private social data that is readily available on Facebook. Our work is based on tracing emerging global threats on Facebook (OSN).We detect the mood of the Facebook users across a group by random walks with the aim of sampling, collecting and monitoring open information through millions of Facebook messages. This will help government agencies to deal with some anti national movement in the nation.*

Index Terms *– crawler*

I. INTRODUCTION

Security techniques are essential for predicting the occurrence of rising global threats through online social networks. The popularity of online social networks (OSNs) is constantly increasing. We selected Facebook to provide programmed searches since it is represented by more than 721 million of active users, the highest number of visitors among online social networks. The security agencies across the world have started to enhance their open social intelligence collection competences. This is done by implementing the mechanism for gathering and sharing publicly accessible information.We believe that open source collection capability can be made better by using biological approach such as Random walks (RWs). Our study has biological basis, but it has direct implications on social networks such as Facebook. Mainly, we studied a set of exploratory algorithms which puts effort to optimize the searching time by means of random movements.

An interactive tool provides us a set of four schemes based on random walk exploration, as well as, it will examine publicly available material posted on Facebook through random walk or optimal social foraging strategy that use the same searching pattern which is used in optimal foraging theory. We are crawling the Facebook and collecting publicly accessibly data for generating the graph of active Facebook users, including path length, clustering,and connected friends. Crawling techniques can be classified into two categories: (a) graph traversal techniques and (b) random walks. In graph traversal techniques, each node in the connected component is traversed exactly once until the completion of the process. Random walks allows re-visiting of nodes and has well known properties for brilliant survey. In our present analysis, we have studied four bio-inspired random walks (Adaptive, Brownian, CRW, Lévy).

Only quarter of Facebook users accept privacy policies. These private policies avoid other users from visiting their friend list. Thus, in view of that there are around 800 million of active users we assume that there is still adequate amount of information for exploring and examining global threats. Additionally, this exploration method enriches the capability to immediately search and perceive key words and character strings across millions of messages, revealing and gathering insights into the combined moods of area or groups abroad.

In this paper we are concerned with the tracking and monitoring the paths generated by the random walks. So, we adjust our exploration in order to follow a resolute path according the stochastic algorithm, so that a directed graph is created during runtime. The amount of active users are enormous and so is the publicly available material on Facebook. So it is essential to develop a method which is able to handle the inbuilt uncertainty and active nature of Facebook. We believe that this kind of examining task could be performed by one of the main concepts introduced by the optimal foraging theory, the random walk.

II. RELATED WORK

There are two categories of Crawling techniques:
1) Graph Traversal

2) Random Walk

Through Random Walk, we could revisit a particular node and could also see its well known properties which were very helpful for the survey. Previous studies have been carried out for collecting information on Facebook.Some of which sample the social graph either in part or totally. Because of the vast data quantity of this social network, the single study able to process this scale of social graph information was presented by Ugander, Karrer, Backstrom and Marlow. An important thing to note here is that this study was sponsored and supported by people working at Facebook.For that reason, the study had complete right of entry to Facebook network.

Earlier studies on open social networks have obtained a homogeneous random user by measures of classic crawling techniques. Though, there were a few drawbacks of using these techniques directly. One of the drawbacks is that we need a lot of resources to work with complete Facebook graph. It is not easy to deal with huge range of mining issues. It is also very complicated to download and deal with terabytes of data without appropriate equipments.

In the previous studies all the calculations were performed on a Hadoop cluster with 2250 machines. The negative aspect to note here is that during crawling the Facebook server responds only with upto a list of 400 friends, further than that the server did not consider the friends list.

There is a privacy setting in Facebook that limits a person from using the information of a person who is not his or her friend. But according to earlier studies there are only few people who use this setting. Therefore, we still have a lot of information for exploring the global threats.

III. STOCHASTIC RULES

Here, we come across a set of four strategies that have a lot to do with "Biological Foraging" . A Very impressive concept here, introduced is the random search, by the optimal foraging theory. A very intricate environment determines the random cause – effect response presented by this type of strategies. tn the current paper , the whole social arena (reachable Howsoever) is considered as a complex environment

Our aim is to explore the data publicly accessible & scrape it with four adapted random models. First one is Brownian motion, totally unbiased, unconnected .Uncorrelated and Unbiased are the terms which have a very peculiar meaning, which needs to be understood properly. This Brownian Model has some features worth noting such as : short distance explorations can be made in much shorter time as compared to explorations over long distances, the random walker actually explores particular region/space rather extensively and chooses new regions to explore blindly before finally being gone again. The random Walker has no past links at all and shows absolutely no tendency to occupy regions that had not been occupied before by it , and hence its track has no contribution in filling up the space uniformly.

The second model presented is the Lévy Walk ,where the most important assumption is that this kind of strategy has many statistical properties such as super diffusion , scale–invariance .Generally Lévy walks are identified by a distribution function which goes as $P(Lj) \sim Lj^{-\mu}$ with $L<\mu<=3$ where the terms have their respective meanings

Lj : Flight Length

~ : Asymptotic limiting behavior

μ: Lévy Index (Exponent of the power Law)

Special cases:

a)when μ>= 3 ; Gaussian Distribution arises as a result of central limit theorem .

b)when μ<=1 ; In this case ,the probability distribution cannot be normalized .

One must know that 'μ' controls a range of correlations in the movement.

The third random walk is CRW (Correlated Random Walk) .Known names such as Raposo Stanley, da Luz & Vishwanathan argue that introducing correlations or say memory effects between every next random walk steps is the best and easiest way to introduce directional persistence into a random walk model. Hence the trajectories made by correlated random walk models have more resemblance to empirical data obtained by direct tracking of animal foraging than those produced by uncorrelated random walks .

Experimental results using Wrapped Cauchy Distribution for the turning angles is

$$\theta = [\, 2 \times arctan \{ \frac{(1-\rho) \times \tan(\pi \times (r-0.5))}{1+\rho} \} \,]$$

Where ρ : Shape parameter ($0 \leq \rho \leq 1$)

r : Uniformly distributed random variable , r € [0,1]

.Note: Directional parameter can be controlled by changing 'ρ' of W(1) . Thus, for $\rho = 0$, a uniform

distribution with two correlations between successive steps is obtained and for $\rho = 1$, a delta distribution at 0^o is obtained, eventually leading to straight line searches.

And Lastly, is an intermittent adaptive random walk , referred to as "adaptive strategy" here. This model basically switches between two models , Brownian one and Lévy walk according to a biological oscillator that is used by Nurzaman et al. And this switch is fundamentally based on environmental charges. Its important to understand here that we calculated

P(t) = exp(-z(t)) with a conditional function where if p(t) =1 , then a Brownian Motion starts, else a Lévy walk is used .

IV. SOCIAL NETWORK ENVIRONMENT

The Facebook social graph can be modeled as G=(V,E), according to Gjoka , Kurant , Butts and Markopoulou where V: set of vertices assuming the role of users , E: set of edges corresponding to the path traced by the social forager .

Here, we have taken interest in tracking and scrutinizing paths generated by the random walks as mentioned in previous section . Thus, according to the stochastic algorithm , we adapt our searches in order to follow a determined path , so that it can help in building a directed graph on runtime .

Now, considering a single user u characterized by a numeric/alphanumeric unique identifier (as assigned by Facebook). Hence u € V and as we are basically interested in collecting information and imitating the paths generated by random walks , we trash the mutual friendship relationship . We here assumed E as a set of edges that comes after the trajectory created by the removal of any of the proposed four random walks. One more assumption, consequently we make is that each edge (u,v) represent a single step dummied by the transition probability of the random walk.

Actually, transition probability is calculated by characterizing the matter publicly available obtained from Facebook , which is straight away associated to a target / search parameter , represented by a key word that the user had defined or sometimes by breaking events (crisis or threats) with this 'target' a search can be made on publicly accessible data on real time ,i.e. bringing users by making a particular request to Facebook which is 'GET' ,through the Graph API Provided by the OSN . API helps us have a simple ,constant view of the Facebook social graph , uniformly representing things in the graph (e.g. people , events, photos , pages ,etc) and how they are connected (e.g. photo tags ,shared content etc.)

Henceforth, we can ask the ID of the users with the intention to store them in a dynamic list . One must pay attention to the fact that the initial order of users gathered by the GET requests only and only depends on the servers exclusively managed by Facebook.

Then, we begin to remove all copied nodes from A & we choose one specific random walk for picking blindly, users in A. Consequently, all the users are copied to a different matrix B and they are placed in the order they were chosen by any of the four random strategies.

V. DESIGNING THE PATHS IN THE SOCIAL GRAPH

In this paper a conceptual analogy is being determined between an ecological environment considering the path followed by social foragers as the core of this effort . Here, various layouts are considered which are provided by the Java Universal Network which actually is a software collection/library that helps one with extendible and common language for the purpose of analysis, modelling and data can be viewed as a graph or network .JUNG library provides algorithm which generate layouts such as Fruchterman-Reingold Layout, Circle Layout, Spring Layout, Spring2 Layout, Kamada-Kawai Layout , etc .

Basically, nodes are the framework of layouts, assuming the role of active users in Facebook . So that one can use their 'Username' as an ID and access pages. All publicly accessible data of a user can be accessed through graph API. Say, in our interactive tool , any node is picked , this particular action will lead to returning of all the publicly accessible data about the respective user in your default internet browser by opening a window . Also, one needs an explicit permission from that user to get additional information about that particular user .An " access token " is needed for the Facebook user at a higher level .After it has been obtained one can carry out authorized requests for that user by incorporating /including the one's Graph API requests, the "access token " .

VI. EXPERIMENTS ON REAL TIME DATA

Here, in this study , a set of four random walks as mentioned in previous section has been employed as sampling methods for checking material scraped from Facebook . This survey begins by using a searching parameter , a keyword that is provided by user . According to a general power-law distribution , our social foragers does a random walk by going after a search pattern, it also includes contacting the Facebook

servers by the social foragers which helps in providing credentials required for the authentication , making use of token . The messages currently posted on the timeline of the current node /user are explored by the forger, once logged in. This process is performed synchronizing our Random walks with the 'in sequence' content acquired from facebook search engine.

One must note ,that data investigation is not at all biased anyway as we are only interested in exploring making use of a searching mechanism .

The Facebook company mandates that ,users are allowed to access private data of any other user only and only if they belong to their friendship network .Thus , obviously we considered publicly accessible data of those users with their privacy settings , as granted by facebook , as 'default' . Also , there is boundary which doesn't allow fetching information beyond a threshold. However, this study is entirely based on recent emerging threats (global) , so only hourly or daily exploration is needed.

A java program was written with the intention of retrieving messages recently posted by making use of a keyword. Facebook provided API helped carrying out crawling process through it. And all these experiments were carried out on a machine running windows 7 with 3.33 GHz speed on an Intel Core Duo processor.

VII. RESULTS

An interface is created in the system through which a user can enter a key word and all the possible results are displayed creating a directed graph generated by JUNG. We intend to justify our assumptions about a suitable way of monitoring using random walks to keep a tab on emergent global threats in Facebook .

Hence, 10 keywords have proposed to test the made assumptions. The keywords are : MH370, Crimea, Kejriwal ,Obama, china, naxalite etc , taken from breaking news . On the other hand , we must consider the fact that sentiments and mood analysis have become a topic of prime focus in defence and intelligence communities , so we also track words such as : happy, sad , angry, riots etc.

The complete list is shown in table :

From Breaking News	For Mood Analysis
MH370 Crimea Kejriwal Obama China Naxalite	Happy Sad Angry Riots

VII. CONCLUSION

In this paper we have presented animal foraging techniques that are used for exploring publicly available information. We have studied four bio inspired random walk (Adaptive, Lévy, CWR and Brownian). This study is based on the first analysis comparing animal movement pattern with the model of searching and sampling to find the threats that are raised in real time.

The study in this paper is biologically aggravated, but it has straight inference in open social network such as Facebook. The authors who carried out the study have provided a collection of probing algorithms that make an effort to optimize the searching time with the help of random movements statistically generated.

Through this paper we have seen that random walks can be used for tracing intimidation in OSN's. The future work in this is to find more searching strategies and to create more random walks for improving and optimizing the power to search in more OSN's.

REFERENCES

[1] L. Cutillo, R. Molva, and M. Onen, "Analysis of privacy in online social networks from the graph theory perspective," in Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE. IEEE, 2011, pp. 1–5.

[2] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," in Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011, pp. 635–644.

[3] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," Arxiv preprint arXiv:1111.4503, 2011.

[4] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou, "Walking in facebook: A case study of unbiased sampling of osns," in INFOCOM, 2010 Proceedings IEEE. IEEE, 2010, pp. 1–9.

[5] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Crawling facebook for social network analysis purposes," Arxiv preprint arXiv:1105.6307, 2011.

[6] J. Keller, "How The CIA Uses Social Media to Track How People Feel,"[Online] Available: http://www.theatlantic.com/technology/archive/2011/11/how-the-cia-uses-social-media-to-track-how people-feel/247923/.2012 4th Computer Science and Electronic Engineering Conference (CEEC) University of Essex, UK46