Visualization of Evolving Large Scale Ego-Networks

Rui Sarmento, Mário Cordeiro and João Gama
LIAAD-INESC TEC
Rua Dr. Roberto Frias, s/n
Porto, Portugal 4200-465
mail@ruisarmento.com, pro11001@fe.up.pt, jgama@fep.up.pt

ABSTRACT

Large scale social networks streaming and visualization has been a hot topic in recent research. Researchers strive to achieve efficient streaming methods and to be able to gather knowledge from the results. Moreover treating the data as a continuous real time flow is a demand for immediate response to events in daily life. Our contribution is to treat the data as a continuous stream and represent it by streaming the egocentric networks (Ego-Networks) for particular nodes. We propose a non-standard node forgetting factor in the representation of the network data stream. Thus, this representation is sensible to recent events in users networks and less sensible for the past node events. The aim of these techniques is the visualization of large scale Ego-Networks from telecommunications social networks with power law distributions.

Keywords

Social Network Stream Mining; Real Time Applications; Data Stream Visualizations; Telecommunication Networks; Ego-Networks

1. MOTIVATION

Visualization of large networks is known to be a hard problem to solve with typical hardware or software. Networks with more than a few thousands nodes and edges can not be computed in limited standard systems. The software and the user himself are the main constraints in visualization tasks of large networks [5]. Even if the software is capable of outputting a network of millions of nodes on the screen it is a very hard task for the user to grasp valuable information from the visual outcome or from its analysis. In this paper we discuss and propose a new way of outputting the data as a network data stream to help the observer visualize the network and enable knowledge acquisition from the output. We present a summarization method for large scale online network streaming focusing on any specific node. Moreover, we assemble existing and novel algorithms for visualization or analysis to get a more efficient method. The results were obtained by simulation of data streaming, originated from databases. All experiences were executed with an ordinary commodity machine.

http://dx.doi.org/10.1145/2695664.2695960

2. RELATED WORK

Several studies have already addressed Ego-Networks. This field covers and relates with many varied subjects, from biology to sociological and criminal networks. The section introduces related work with a more generic approach including an overview of research on social networks.

2.1 Ego-Networks

In [4], a throughout exposition about Social Networks is made and a full chapter is dedicated to Ego-Networks. Hanneman et al. define "Ego" as an individual "focal" node in a network. "Neighborhood" defines the boundaries of ego networks and includes all the direct connections and egos that tie with an ego. DeJordy et al. [2], introduce the network perspective and the differences between socio-centric and ego-centric analysis. The ego-centric approach fits studies about phenomena or entities across different networks. The socio-centric approach is more suitable for studying interaction within a defined network. Wasserman et al. provide a complete study about social networks with several models in [6]. Some important studies address the social structure of competition. For Burt et al. [1], social structure of competition addresses the consequences of voids in relational and resource networks. Competitive behavior can be understood in terms of player access to "holes" in the social structure of the competitive arena. Those "structural holes" create entrepreneurial opportunities for information access, timing, referrals and control. Ego-Networks analysis provides an answer to this sensible information or properties that are are also studied in the case study section of this document.

3. STREAMING ALGORITHM

In this section we describe the algorithm used for streaming. For the Ego-Networks streaming representation of the data we used the landmark window implementation [3], focusing on the ego node's 1^{st} and 2^{nd} order connections. This algorithm represents an alternative to full networks SNA model. Fig. 1 shows the example Ego-Network for the network node in light green colour at the centre of the picture, its direct connections and its 2^{nd} order connections.

The developed application enables the visualization of the events centred on some specific network node, instead of the evolution in time of the full network events. The input is a

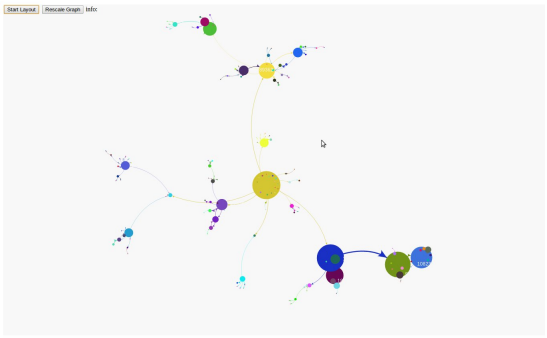

Figure 1: Visualization of an Ego-Network

start date and hour, as well as the end date and hour in the landmark window. The algorithm returns the visualization of the evolving Ego-Network over time. New connections arising from the central node or their 1st order connections are plotted in the screen.

3.1 Ego-Networks with forgetting factor

Ego-Networks with node forgetting factor algorithm, using streaming simulation, implies the update of a node structure with a *forgetting factor* variable per node. This variable value is the same for each new node represented on the graph. After some estimated streaming time period, forgetting factor values are updated for all nodes currently present in the structure. Those nodes with lower than the threshold forgetting factor are deleted from the graph along with their direct outgoing connections. The following expression presents the forgetting factor update that was used

$$f_n(t) = (factor)^{p_n}$$

where $factor$ is the selected initial forgetting factor, with $0 < factor < 1$, and p_n is the number of update periods for node n. p_n will be set to zero every time node n establishes a new outgoing connection. Thus, nodes that have recent related events are kept using the data timestamps.

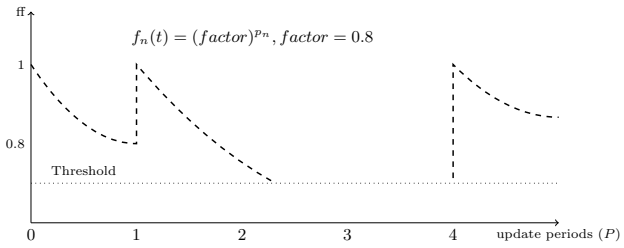

Figure 2: Evolving inclusion or exclusion of a given node based on the forgetting factor function. P_n is the number of update periods considering that node n did not receive any connection

Fig. 2 describes the behaviour of the forgetting factor function for a specific node. In this example, the represented node establishes connections for the update periods 0, 1 and 4, meaning that for these update periods the number of periods without being updated will be zero ($P_n = 0$), and therefore $f_n(t) = (0.8)^0 = 1$. For the other update periods, ex.: 3, the value of the function will be $f_n(t) = (0.8)^2 = 0.64$, with $P_n = 2$. This means that at the update period 3, the node did not established a connection for two update periods. The figure also shows the threshold value for considering, or not, a particular node. When the function crosses bellow the threshold for a particular update period, the node is removed from the graph. In the example, right after the update period 2, the node will be removed from the network and will only be considered again in update period 4, when the node establishes a new connection.

4. CASE STUDY

In this case study, the proposed forgetting factor method was tested in large scale telecommunications Ego-Networks. The aim was to find out if the method results were representative of the original data Ego-Network as the data streaming and forgetting factor value evolved over time. Thus, we used our novel algorithms either with and without forgetting factor. For these tests we discarded any events related to voicemail numbers which biased these studies.

4.1 Data Description

The used Call Detail Records (CDR) log files were retrieved from equipment distributed in different geographic locations. The network data has an average of 10 million calls (edges in the social network) per day. The phone numbers were changed to different identifiers to preserve users anonymity. A call between A and B phones is represented as an edge in the social network. Because some individuals receive and make more than one call, the full networks has an average of 6 million of unique users/nodes per day. The dataset contains anonymous data for 135 days. For each edge/call, timestamp information shows the date and hour of the beginning of the call. The number of calls per second varies from around 10 at mid-night and reaches its peak at mid-day with 280.

4.2 Ego-Networks with forgetting factor

Fig. 3 illustrates the test of parameters with the same anonymous number's Ego-Network presented in Fig. 1. The *update time period* was set to 3 hours and the *forgetting factor threshold* set to 0.6. The *initial forgetting factor* value was 0.95.

For the same period of streaming Fig. 3 presents much less connections in the graph, meaning that some of the connections of the central node were deleted as an outcome of a forgetting factor lower than the threshold value of 0.6.

4.2.1 Node and Edges counter variation with time

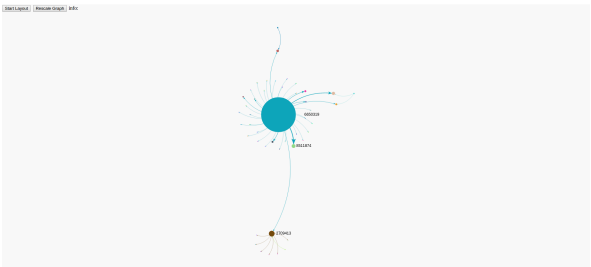

Figure 3: Ego-Network with forgetting factor

After the previous experiment we randomly selected three phone/nodes in the set of approximately 10000000. This was made for one week period of the available data. With these nodes we run the experiments with 5 different update periods, 3 different thresholds and 3 different forgetting factor values. The used values for the experiments were 3, 6, 12, 24 and 48 hours for the *update time period*. The selected values for the *initial forgetting factor* were 0.65, 0.8 and 0.95. The *forgetting factor threshold* values were 0.1, 0.3 and 0.6.

We then studied the variation of the counters for nodes and edges. The variation regarded time and comparison between Ego-Network with and without forgetting factor for the same update periods. It only considered 1st order outgoing connections for the experiments. Thus, the focus is more on active node events and less on the passive ones.

The following figures represent the variation when the *forgetting factor* is 0.95 and the *forgetting factor threshold* is 0.6 and by changing only the *update time period* parameter.

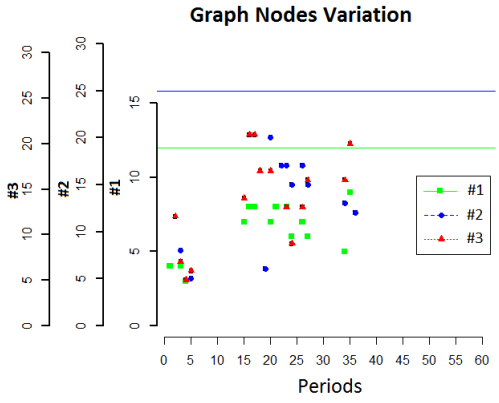

Figure 4: Ego-Network nodes counter

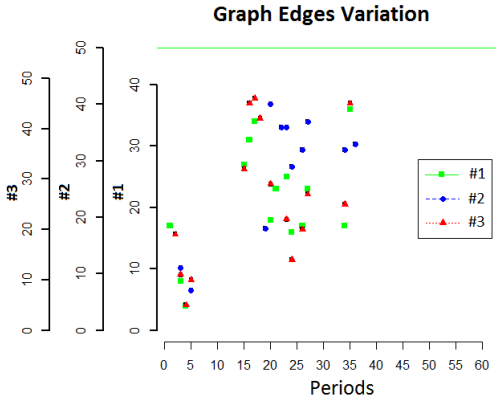

Figure 5: Ego-Network edges counter

It is known that the values of both counters increase with time for the original Ego-Network without forgetting factor. In Fig.4 and Fig. 5, the horizontal lines represent each node network's maximum number of nodes and edges in graph. With the forgetting factor version, in Fig. 4 and 5 with update period of 21600 seconds, there is evidence, for the majority of the three selected numbers, the same counters are lower and with tendency for stability. Moreover, the maximum number of nodes and edges are much lower than the original network maximum values for both counters. Thus, some horizontal lines are off the scale and therefore not visible in these figures. It proves this method of streaming implies memory saving characteristics by discarding older events in the streaming.

5. CONCLUSIONS

In this short paper we propose a new type of application for large scale telecommunications networks visualization, streaming and analysis. With the use of data time stamps we approach the data with a streaming point of view to visualize samples of data. Thus, it is both comprehensible to the user and also enables knowledge extraction from the visual output.

Variations of the Landmark Window streaming algorithm were developed to output Ego centred networks and ultimately, by using forgetting factor in the Ego-Network output, we propose a novel streaming method for this type of social network analysis. The weight and importance of recent events is higher for the majority of the test cases. We consider and prove this to be an effective visualization method for Ego-Networks SNA.

Finally, we conclude that this method for evolving networks visualization and analysis is a light method to visualize massive Ego-Networks. Thus, the simulation of a data stream and the visualization results very close to the node-link level can be achieved using an ordinary commodity machine.

Future work will engage other enhancements to real time data streaming, leveraging telecommunication systems and enabling the visualization of real time evolving Ego-Networks.

6. ACKNOWLEDGMENTS

This work was supported by Sibila and Smartgrids research projects (NORTE-07-0124-FEDER-000056/59) , financed by North Portugal Regional Operational Programme (ON.2 O Novo Norte), under the National Strategic Reference Framework (NSRF), through the Development Fund (ERDF), also by national funds, through Fundação para a Ciência e a Tecnologia (FCT), by European Commission through the project MAESTRA (Grant number ICT-2013-612944) and the project number 18450 through the "SI I&DT Individual" program by QREN and delivered to WeDo Business Assurance.

7. REFERENCES

[1] Burt, R. S. *Structural holes: The social structure of competition.* Harvard University Press, 1992.

[2] DeJordy, R., and Halgin, D. Introduction into ego network analysis, 2009.

[3] Gama, J. *Knowledge Discovery from Data Streams*, 1st ed. Chapman & Hall/CRC, 2010.

[4] Hanneman, R., and Riddle, M. *Introduction to Social Network Methods.* University of California, 2005.

[5] Rafiei, D., and Curial, S. Effectively visualizing large networks through sampling. *Visualization Conference, IEEE 0 (2005), 48+.*

[6] Wasserman, S., and Faust, K. Social Network Analysis: Methods and Applications. *Cambridge University Press, 1994.*