Sampling Evolving Ego-Networks with Forgetting Factor

Shazia Tabassum
LIAAD, Inesctec
University of Porto
Porto, Portugal
Email: up201402360@fe.up.pt

João Gama
LIAAD, Inesctec
University of Porto
Porto, Portugal
Email: jgama@fep.up.pt

***Abstract*—Dynamically evolving networks get humongous in no time. Usually, sampling techniques are used to create representative specimens of such large scale socio-centric temporal networks. Likewise, the size of ego-networks gets larger over a period of evolution. Which is why, there is a need to sample ego-centric networks while maintaining the importance and efficiency of the ego. In this paper, we present a novel method to sample ego-networks as they evolve, while maintaining the freshness of the ego-network, with the latest ties and most stronger relationships from past, based on an attenuation factor. We made use of an exhaustive list of node level and graph level metrics to evaluate and compare the samples with the original network. Our experiments show that the proposed method maintains most active and recent nodes. It also preserves the strength of ties between them. We find that our method decreases the redundancy while maintaining the efficiency of network. We also analyzed the evolution of an anonymised phone calls ego-network over a period of 31 days.**

1. Introduction

Ego-centric network focuses on the relationships of a single node in a virtual interaction network graph. An ego can represent an individual, entity, object or organization. All the other nodes in the network of an ego are alters. An ego centric network maps the relationships of an ego with alters and also between themselves. In the recent work [1] by Google.com, the authors argue that it is possible to address important graph mining tasks by analyzing the ego-nets of a social network and performing independent computations on them. In [2] Wellman describes an ego-network as a personal network. The author explains that the importance of local ties becomes apparent by redefining the composition of personal community networks in terms of the number of contacts (interactions) that egos have with the active members of the networks instead of the traditional procedure of counting the number of ties(relationships).

Networks representing real world social structures are usually temporal and evolving. The rapidly changing and evolving structure of these graphs, calls for an exigence of latest and up to date results. Processing the real-time network stream as it arrives, is one of the best solutions for the above problem. Therefore, we use a streaming ego-network approach over a telecommunications' call graph stream of temporal edge/calls'. In [3] the authors discuss methods for analyzing large-scale call networks. Call networks are one of the largest and fastest networks with evolving and inconsistent tie strengths. Therefore, we need a method for real-time sampling, while preserving the importance of tie strengths for the applicability of such networks.

Capturing the ego-networks of high velocity streaming graphs over a period of time is highly infeasible as it can reach over millions of redundant edges for an active user in few days. If we only sample nodes, we miss the tie strengths of alters with egos, as previously investigated in our work of [4]. Therefore, we base our sampling method using edges. Now the obvious question is, how do we capture the ego-network of an evolving multi-graph stream over time with least possible edges, while preserving the structure, properties and efficiency of an ego-network? For which, we propose an ego-network sampling method using a forgetting factor. The proposed method is suitable for dynamically evolving multi-graphs. We use our method over a real world temporal stream of edges/calls to generate a sample stream in real time. Our results show that the proposed method preserves tie strengths in the ego-networks. We also show that our method decreases redundancy in the network while preserving the importance of ego. We measure the importance and efficiency of network using some socio-metrics. We evaluate our method by comparing the samples generated by varying parametric values, with the original ego-network. The proposed method can also be implemented over a socio-centric network.

2. Related Work

The concept of ego-networks was discussed by L.C.Freeman in [5], where he described an ego-network as a social network, built around a particular social unit called ego. In [2] Wellman discusses the importance of local ties in personal networks. In [6] Burt studied the affects, gaps and relationships between the neighbourhood of a node, referring them as structural holes. He also introduced metrics to evaluate an efficient-effective network which strives to

optimize structural holes in order to maximize information benefits.

[1] uses ego-net community mining for suggesting friends in a social network, based on the co-occurrences of two nodes in different ego-network communities. In [7] the authors compared the socio-centric and ego-centric approaches for a social network data collection procedure.

[8] proposed an ego-centric network sampling approach for viral marketing applications. The authors employed a variation of forest fire algorithm for sampling ego-network. They compared the degree and clustering coefficient distributions of sampled ego-networks with the original ego-network. However, to the best of our knowledge there are no methods for sampling ego-networks in real-time with evolving samples. In this work, we introduce an edge based sampling method with forgetting factor over an evolving ego-network stream of temporal edges.

3. Sampling Ego Network with Forgetting Factor

In this method, we sample edges from a stream of temporal network. As sampling edges preserves the structure of network, while sampling nodes we need to acquire the adjacent edges and also the adjacent nodes. We start by building the ego-network of a specific ego and begin to scrape together all the adjacent ties to the ego and their adjacent ties. We do this by using a set for storing adjacent nodes. For every recurring edge, we increment the edge weight of the corresponding edge by maintaining them in a hash table. We impose a forgetting factor over each edge weight, following fixed time period t. In our experiments, we use a $t = 1$ day. This means we apply the forgetting factor over the ego-network as soon as the stream enters a new day, i.e we forget the old edge weight $w(t-1)$, by some fixed percentage defined by the forgetting factor and sum up with the latest weight of edge w_t (from the edges arrived in time period t). The forgetting factor is given by two parameters, an attenuation factor α and a threshold θ. Where $0 < \alpha < 1$ and also $0 < \theta < 1$. After every t the tie strength between two nodes is given by the function $w(t)$ in equation 1.

$$w(t) = w_t + (1 - \alpha) \times w(t-1) \tag{1}$$

After every successive time period, we decrease the edge weight by α and consequently remove the alter/alters adjacent to the corresponding edge as the edge weight decreases than the threshold value θ. When α=1 we have a maximum forgetting i.e we forget the whole network except the network of current day. When $\alpha = 0$ we get the original network. If the removed edge corresponds to an alter adjacent to the ego, we remove the adjacent edge and the alter, and all the second level alters adjacent to the alter itself, if the above condition is satisfied. If we forget a second level edge, not having a direct connection to ego then we only forget the corresponding node. Following this strategy, we can have most active alters in the ego-network at the end of each day.

4. Metrics for Evaluating Ego Networks

In this section we discuss an extensive list of graph metrics used in this scenario to measure the structural, topological and behavioral properties of the ego-networks. We exploit these properties at graph level and node level.

4.1. Graph level metrics

In this work, we studied the properties of ego-network graphs using average degree, average weighted degree, density, diameter and average path length.

Additionally for evaluating evolving samples using our proposed method, we compared the degree distributions of the samples at the end of 31 days with the original network using kolmogorov-Smirnov test. We use the D-statistics from the test and also p-values to evaluate our null hypothesis (H_0) that our sampled ego-networks follow the same distribution as the original ego-network. The degree distributions of the networks is obtained by counting the frequency of each degree d in the network. The frequency of each degree d is given by the number of nodes with degree d in the network snapshots at the end of 31 days.

We compared the effective size and efficiency of samples with that of ego-network using ego metrics introduced by Burt in [6]. Effective size of the ego-network (ES) is the number of alters that an ego has, minus the average number of ties that each alter has to other alters. In the simplest form, for an undirected ego-network of radius 1, the effective size can be given by the eq.2. Efficiency (EF) of an ego-network is the proportion of ego's ties to its neighborhood that are "non-redundant." Efficiency is the normalized form of effective network size (eq.3). Therefore, it is a good measure for comparing ego-networks of different sizes.

$$ES = n_a - \frac{\sum_{a=1}^{n_a}(d_a - 1)}{n_a} \tag{2}$$

$$EF = \frac{ES}{n_a} \tag{3}$$

where n_a is the number of alters in the ego-network and d_a is the degree of an alter a.

4.2. Node level metrics

We make use of a bunch of centrality metrics to study the importance of ego and also compare the position of ego in the original network with the sample networks. Measures include degree, weighted degree, closeness (CC), and eigen vector centralities (EVC). We also explored the eccentricity and clustering coefficient of the ego.

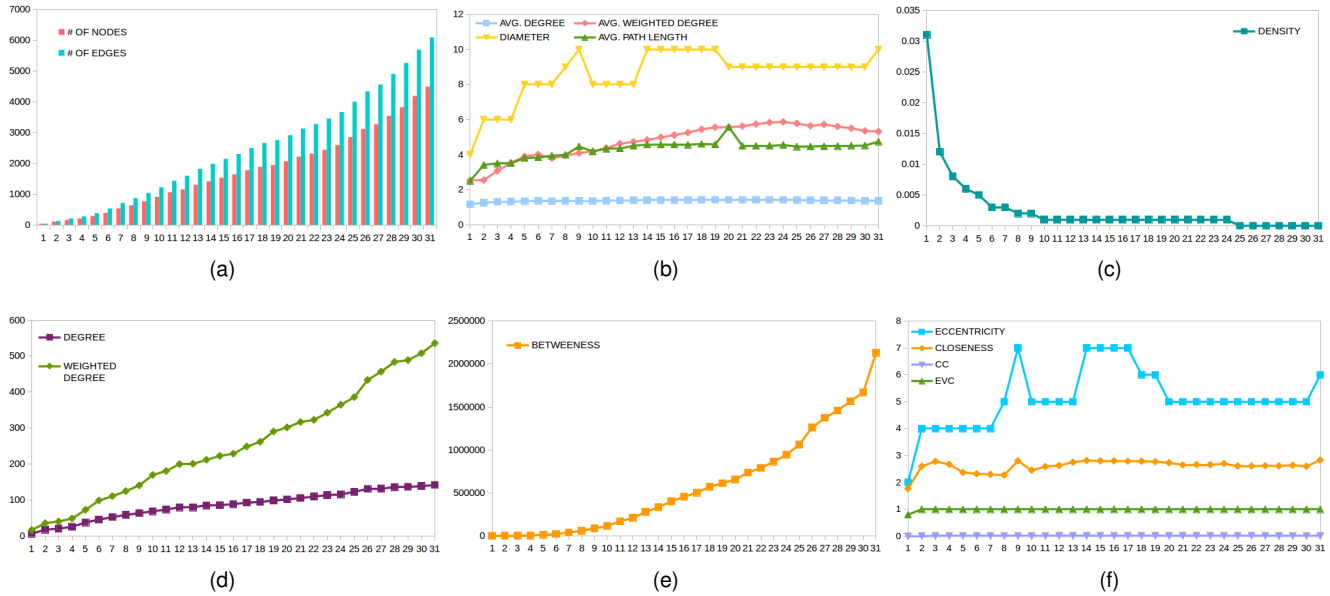

Figure 1. Evolution of a call ego-network over a month

5. Evaluation Methodology

In order to evaluate our method SEFF discussed in section 3, we applied it over a real world streaming call Graph G of 31 days by randomly choosing an ego e and generating a sample stream of depth $d = 2$ at any point of flow. This was done by generating six real time sample streams, where each sample stream S_i is generated by different combinations of $\alpha \in \{0.9, 0.8, 0.7, 0.5\}$ and $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ discussed in section 3. For investigating the above sample streams, we captured their snapshots at the end of 31 days each. Each snapshot $S_i^{31} \subset G$. Beforehand, we took a snapshot G_e^{31} of original ego-network stream G_e of e (where $d(G_e) = 2$) at the end of 31 days from the socio-centric call graph G. Each sample graph $S_i^{31} \subset G_e^{31}$. We then compared the conclusive sample snapshots S_i^{31} where $1 \leq i \leq 6$, with the original ego-network snapshot G_e^{31} by using metrics discussed in section 4. Conclusively, we derive some conclusions about the structural properties and behavioral properties preserved by the sample networks.

Additionally, we also analysed the evolution of streaming ego-network for 31 days by grabbing snapshots G_e^n where $1 \leq n \leq 31$, of original ego-network at the end of each day stream. Eventually, we perform a piecemeal structural analysis over the augmented samples by employing measures discussed in section 4.

6. Case Study

Telecommunications' call graphs are very rich in depicting real world social relations as phone calls represent edges between real world users. For this we made use of an anonymised call stream of 31 days available from a service provider. The network data stream was generated a speed of 10 to 280 calls per second around mid-night and mid-day. On an aggregated scale we have 400 million calls made by 12 million subscribers.

We considered call networks as a special application scenario for employing our method as these networks are 1. High speed streaming and evolving temporal graphs. 2. Multi-graphs with more than one edge between two users, representing the strength of their relationship unlike a social network based on friendship and, follower and followee relations, where there is a single binomial relation between two nodes. However, the proposed method can be applied to networks with binomial relationships as it forgets edges and eventually forgets nodes. SEFF method is also appropriate for sampling weighted networks.

7. Experimental Evaluation

To analyze the evolution of a call ego-network, we selected an arbitrary user "ego" from the real world call/edge stream described in section 6 and built its ego-network in a streaming fashion as described for samples in section 3. Then we grabbed the snapshots of ego-network at the end of each day and measured their graph properties discussed in section 4. Figure 1 depicts the evolution of ego-network per day. Now we have the snapshots of original ego-network for 31 days. Using the same ego we generate six evolving sample streams as discussed in section 5. Six different combinations of α and θ corresponding to six different samples are shown in figure 2. The figure also plots the values of computed metrics over the conclusive sampled ego-networks and the original ego-network.

Figure 2(a) shows the number of nodes and the number of edges in the above described ego-networks. We observe

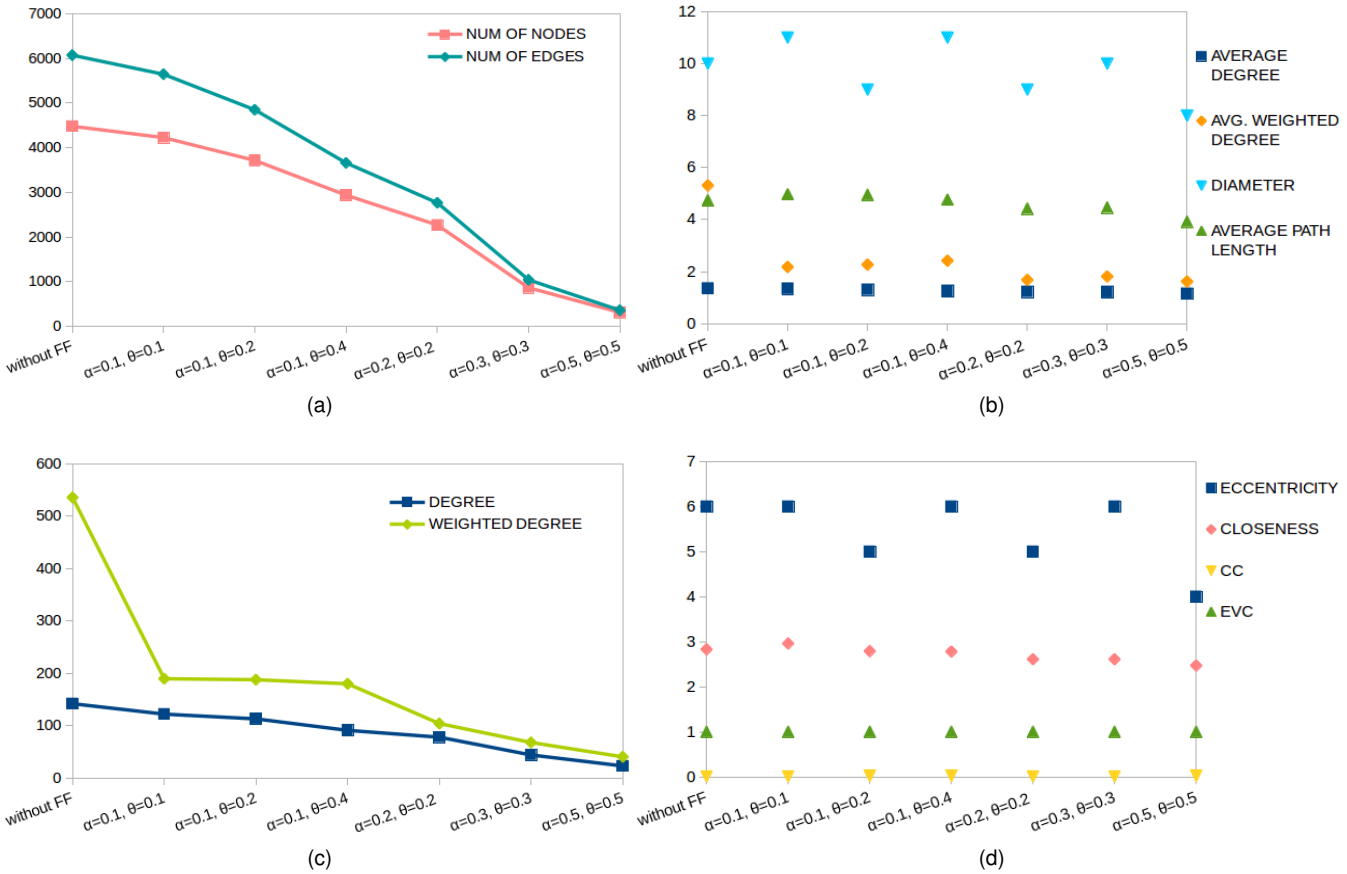

Figure 2. Metrics over ego-networks with and without forgetting factor

that the number of nodes gradually decrease with the increasing forgetting factor. For an attenuation value of 0.5 and threshold value of 0.5 we forget 50% of the edges per day, between two adjacent nodes. As a result, we have most active nodes as we increase forgetting. We also observe that, the number of edges decrease in greater proportion than the number of nodes, Almost reaching equal, for the highest forgetting factor in the illustration. This exhibits that the proposed SEFF method decreases redundant edges.

Figure 2(b) depicts metrics over the ego-networks. The diameter of the graphs varies with the inclusion and removal of the connecting nodes from the ego-network. It depends on the network of ego selected. Average degree and the average path length decreases with the increasing forgetting, this shows that the networks shrink with increased forgetting. SEFF method has a noticeable effect over the weighted degree of graphs.

The degree and weighted degree of ego are plotted in figure 2(c). Both the values decreased with the increased forgetting, while the drop in weighted degree is higher. This suggests that when we increased forgetting, we actually decreased the tie strengths but relatively maintained the ties. In Figure 2(d) we see that the eccentricity has a similar effect of diameter in the ego-network graphs. This corresponds to the conceptual relation between diameter and eccentricity. Closeness of the ego with alters also decreased gradually with the increased forgetting factor. The clustering coefficient of ego is too low to compare. The eigen vector centrality portrays the important node in the network. SEFF preserves the importance of ego along side forgetting.

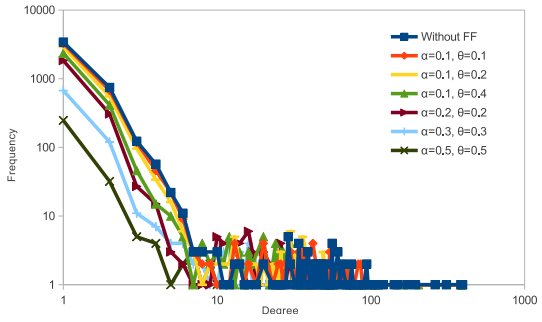

Figure 3. Degree distributions of ego-networks at the end of 31 days with and without forgetting factor

We also compare the degree distributions of the original ego-network with the samples generated by using SEFF method at the end of 31 days. We applied Kol-

TABLE 1. COMPARISON OF DEGREE DISTRIBUTIONS USING KS-TEST

Samples	α=0.1, θ=0.1	α=0.1, θ=0.2	α=0.1, θ=0.4	α=0.2, θ=0.2	α=0.3, θ=0.3	α=0.5, θ=0.5
D-stat	0.146	0.138	0.173	0.146	0.191	0.096
p-value	0.114	0.124	0.065	0.182	0.105	0.724

mogorov–Smirnov test to compare the degree distributions of the samples with the original network. The D-statistics and P-values of tests are given in the table 1. The p-values are computed using exact method. The significance level used for the comparisons is 5%, ie α=0.5. The results show that all the sampled distributions follow the distribution of original graph. We also observe that the value of θ has a greater impact on the similarity of distributions, than α in the SEFF method. We can see the pictorial representation of the degree distributions of the original graph and sample graphs in fig 3

Figure 4 illustrates the effective size and efficiency of the ego-networks. Decrease in effective size shows decreased redundancy. Efficiency of the network indicates the impact of ego in the network. In the given figure we can observe that the efficiency of the network is maintained through out the samples using SEFF while decrease in the effective size.

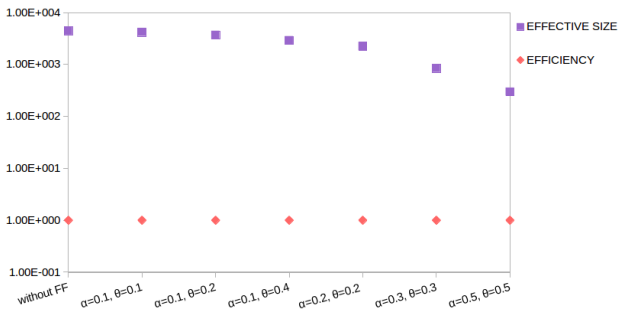

Figure 4. Efficiency and effective size of ego-networks

8. Conclusion

We presented a novel method of sampling an evolving ego-network with forgetting factor. We discussed our method in terms of a streaming network. However, it can also be applied over an aggregated network with time stamps. We carried out experiments by using a real world calls' network, which is a multi graph finding it more appropriate scenario for SEFF. We also examine the evolution of an ego-network for a period of 31 days. To compare the samples and analyze the evolution of ego-networks, we exploited a number of metrics. We prove that our method maintains the latest and active nodes in a network and also preserves the properties of it.

Acknowledgments

Author acknowledges the support of the European Commission through the project MAESTRA (Grant Number ICT-750 2013-612944), FCT (Fundação para a Ciência e a Tecnologia) within project UID/EEA/50014/2013 and also thank WeDo Business for providing the data.

References

[1] A. Epasto, S. Lattanzi, V. Mirrokni, I. O. Sebe, A. Taei, and S. Verma, "Ego-net community mining applied to friend suggestion," *Proceedings of the VLDB Endowment*, vol. 9, no. 4, pp. 324–335, 2015.

[2] B. Wellman, "Are personal communities local? a dumptarian reconsideration," *Social networks*, vol. 18, no. 4, pp. 347–354, 1996.

[3] R. Sarmento, M. Oliveira, M. Cordeiro, S. Tabassum, and J. Gama, "Social network analysis of streaming call graphs," in *Big Data Analysis: New Algorithms for a New Society*. Springer, 2015, vol. 16, pp. 239–261.

[4] S. Tabassum and J. Gama, "Sampling massive streaming call graphs," in *ACM Symposium on Advanced Computing*, 2016, p. In Press.

[5] L. C. Freeman, "Centered graphs and the structure of ego networks," *Mathematical Social Sciences*, vol. 3, no. 3, pp. 291–304, 1982.

[6] R. S. Burt, *Structural holes: The social structure of competition*. Harvard university press, 2009.

[7] K. K. Chung, L. Hossain, and J. Davis, "Exploring sociocentric and egocentric approaches for social network analysis," in *Proceedings of the 2nd international conference on knowledge management in Asia Pacific*, 2005, pp. 1–8.

[8] H. H. Ma, S. Gustafson, A. Moitra, and D. Bracewell, "Ego-centric network sampling in viral marketing applications," in *Mining and Analyzing Social Networks*. Springer, 2010, pp. 35–51.