International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 07 | July 2018 www.irjet.net p-ISSN: 2395-0072

Efficient online summarization of large scale dynamic networks in social network

Mrs. Rupali Molawade[1], Prof. Deepa Parasar[2], Prof. Sujata Bhairnallykar[3]

[1]*ME Student, Computer Engineering Department, Saraswati College of Engineering, Kharghar,Navi Mumbai*
[2,3]*Associate Professor, Computer Engineering Department, Saraswati College of Engineering, Kharghar,Navi*

Abstract - *Social media forms a central domain for the production and dissemination of real-time information .Information diffusion is fundamental process taking place in social network. While it is often possible to directly observe when nodes become infected or publish the information, observing individual transmissions (who infects whom, or who influences whom) is typically very difficult. Furthermore, in many applications, the underlying network over which the diffusions and propagations spread is actually unobserved. A key tool in this regard is data summarization. However, few existing studies aim to summarize networks for interesting dynamic patterns. Dynamic networks raise new challenges not found in static settings, including time sensitivity, online interestingness evaluation, and summary traceability, which render existing techniques inadequate. We propose dynamic network summarization to summarize dynamic networks with millions of nodes by only capturing the few most interesting nodes or edges over time. Based on the concepts of diffusion radius and scope, we define interestingness measures for dynamic networks, and we propose OSNet, an online summarization framework for dynamic networks. Efficient algorithms are included in OSNet.*

Key Words: **Information Diffusion, Dynamic Network, Interestingness, Online summarization**.

1. INTRODUCTION

Diffusion of information, spread of rumours and infectious diseases are all instances of stochastic processes that occur over the edges of an underlying network [10].Many times networks over which contagions spread are unobserved, and such networks are often dynamic and change over time. Driven by the ever-increasing sizes of real world networks, the problem of network summarization has never been more important [14]. While most existing studies consider the summarization of static networks according to criteria such as compression ratio, network representation, minimum loss, and visualization friendliness recent developments in social network mining and analysis location-based services and bioinformatics have given prominence to the study of a new kind of dynamic network that captures progressive information diffusion processes in an underlying network. An information diffusion [10] process in a network can be represented by a stream of interactions between node instances, namely time-stamped pairs of nodes from the underlying net-work, denoting the information propagation from one node to the other at the time as indicated by the associated time-stamp An example of a diffusion process is the spread of news items among Twitter users by means of the "retweet" functionality [11]. Such a stream describes a dynamic network where a diffusion process grows with each incoming node instance pair as a new interaction in the stream. The summarization task of such a dynamic network [6], which is significantly different from that of a static one, poses new research challenges. The critical difference lies in that, for a dynamic diffusion process, it is valuable to capture each "interesting" development as the process evolves, in an online fashion. This problem, termed as dynamic network summarization (DNS), has a wide range of applications, among which we highlight several as follows. Two useful phenomena observed from social network studies have motivated us to summarize dynamics based on interestingness. First, user interactions on social networks may show the interestingness of time-stamped posts. For example, the release of a new Nintendo DS game as a time-stamped post on Twitter can induce many users' tweets. We consider such a time-stamped post as a node, and the measure of the number of user interactions as node degree has been used for finding hot topics, detecting burst events, and mining influential users. Second, social networks are born for social conversations. Interesting topics [7] often provoke long and active conversation chains among a group of users For instance; such a long conversation chain is quite common for interesting questions on Stack overflow, where users have multiple QA discussions on a specific problem to achieve a solution. Multiple involvement of one user in a conversation chain (i.e., node instances) often encourage the participation of other users .Ideally, a conversation chain is traceable such that we can follow conversations from the start to a particular node instance. The traceability is straightforward to demonstrate how users interact, e.g., on Stack overflow, to show a problem and its responses, traceable chains enable us to follow user discussions. The measure of user conversations may have various applications for customer satisfaction survey, user engagement study, and fraud user detection. This study considers the two measures, which intuitively could help us find the summaries where users are influential and actively engaged through particular instances in social activities.

1.1 Research Challenges

We identify the following research challenges in the task of DNS.

1) **Time Sensitivity:** Diffusion processes often represent vast, viral, and unpredictable processes, e.g., breaking news and bursty events As a result, the rate of diffusion can vary drastically over a short period of time .It is a big challenge to respond adaptively to dynamics and to achieve timely summarizations.

2) Online Interestingness Evaluation: A key challenge here is to capture the most interesting nodes and edges in a summarization. Compared with traditional network summarization, interestingness evaluation in DNS assumes an extra degree of difficulty because of the partial view of the network at any given time of evaluation.

3) Summary Traceability: An important goal is to enable a better understanding of the evolution of a diffusion process throughout its life cycle. A good summary should reveal the flow of the dynamics such that interesting developments can be traced.

2. RELATED WORK

While social network analysis has a long history in sociology the rise of the internet and the increasing availability of large datasets have in recent decades sparked interest from other disciplines including (statistical) physics, economics, and computer science. Much of this literature, which has developed quite independently from the existing sociological literature, focuses on the structure and dynamics of large complex networks that may or may not be social networks. The volume by provides a good overview of this literature. The majority of research on online social networks falls within this relatively young tradition of "network science".

Graph compression [1] and simplification mainly focus on generating compact graph representations to simplify storage and manipulation. Much of the work has focused on lossless web graph compression .Web pages with similar adjacency lists are encoded using reference encoding. Most of these studies, however, only focus on reducing the number of bits needed to encode a link, and few compute topological summaries since the compressed representation is not really a graph. An exception is a study that computes graph summaries by grouping web pages based on a combination of their URL patterns and k-means clustering. Lossy topological summarization is another graph compression study.

Compared with these studies, our approach is developed to summarize diffusion processes. Diffusion processes as dynamic graphs do not belong to those special families of graphs (e.g., unlabelled and static trees or planar graphs that have repeated patterns and infrequent change nodes/edges) for which efficient storage compression has been proposed in graph compression literature . As a result, direct adaption of these methods is not possible for online summarization of dynamic networks. Moreover, we use OSNet for city transportation analysis.

As one of the attempts to con-sider time-evolving networks, Liu et al. compress weighted time evolving graphs, which is equivalent to com-pressing a sequence of static graphs. Ferlez et al. propose Time Fall to monitor network evolution that clusters texts in scientific networks and uses MDL to connect clusters. This class of studies are inherently distinct from ours in four aspects: 1) we use general networks and do not have assumptions on text processing; 2) OSNet takes as argument an interaction stream rather than a time-stamped offline net-work; 3) a sequence of time-sliced graphs are not assumed; 4) we aim to summarize diffusion processes. There are also studies on temporal dynamics of social networks, including inferring cascades [1] finding common progression stages in event sequences , predicting cascades . They focus on tasks different from ours.

Many diffusion models[6] are proposed to model information diffusion and adoption, which can be distinguished as explanatory models (e.g., NETINF , NETRATE , INFOPATH) based on complete diffusion data to retrace implicit path from generative probabilistic models and predicting models of cascade unfolding based on historical data . Independent cascade and linear threshold models are two extensively studied graph-based influence diffusion models originally summarized by Kempe et al [13]. The two models are based on the intuition that often decision is correlated with the number of friends. This work does not consider the influence of nodes and makes no assumption on underlying networks. The study takes complete and timely interactions in cascades as information diffusion for the purpose of summarization.

3. PROPOSED WORK

To tackle the DNS problem, we propose OSNet, a framework for online summarization of dynamic networks that aims to produce concise, interestingness-driven summaries that capture the evolution of diffusion processes. Our contribution is summarized as follows. 1) Unlike previous proposals that apply optimization criteria in offline settings, we consider a setting where network summarization occurs online, as the diffusion process evolves. 2) Based on the concepts of propagating radius proRadius and propagating scope proScope, we formalize the problem of characterizing the interesting dynamics of an evolving diffusion process in a traceable manner. 3) We propose OSNet that encompasses online and incremental dynamic network summarization algorithms on a spreading-tree model. In terms of entropy, OSNet archives the best summaries with respect to informativeness. 4) We propose tree-based efficient algorithm.

3.1 PROBLEM STATEMENT

The input to the problem is a stream of time ordered interactions (i.e., diffusion processes) on a network G. A diffusion process on a network G, denoted by *D* (G), is a stream of time-ordered interactions. An interaction x = (δ, u, v, t) ∈ *D* (G) indicates that a specific story is diffused from node u to node v at time t ∈ T. A story is defined by a textual keyword list used to describe an event, such as breaking news in Twitter. The diffusion from *u* to *v* captures that node v receives the story from u. We also say that *u* is an infector of v while *v* is an infectee of *u*. We call time *t* the infection time of node *v*. Note that a diffusion process of a story can be initiated by different nodes that are regarded as seeds or roots. For each interaction *x*, we further define *δ* to be a three-tuple as a canonical identifier, i.e., δ = (storyID, v_r, t_0), where storyID is the identity of the diffusing story, v_r represents the seed node starting the diffusion, and t_0 is the infection time of the infector *u*. The diffusion process from a seed over a time period forms a time-stamped graph, known as a network cascade C where each interaction is a directed edge from the infector to the infectee. The output of a dynamic process as a summary is a subset of interactions with connected node instances of the process.

3.2 Design Details

3.2.1. Frame work

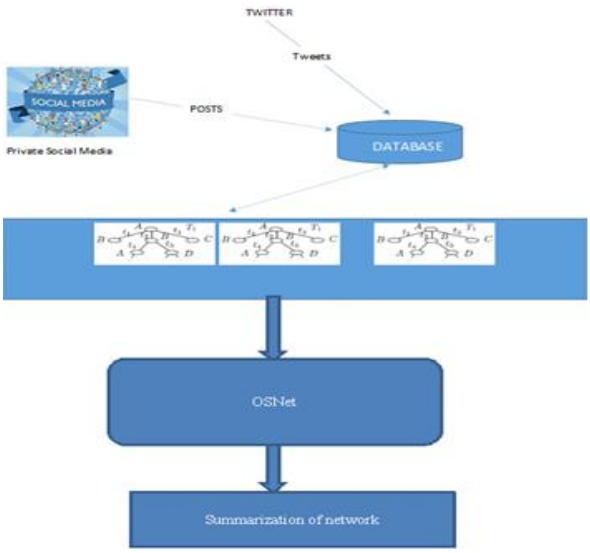

Fig -1: Overview of OSNet Framework

To keep with online behaviour of social network each node (User & POST) Timestamp is associated with each node; same post with diff timestamp is also consider as different node. Only spreading trees satisfying threshold are consider for OSNet Analysis. Even POST is represented by tree of likeness, OSNet will treat tree as Graph for Summary calculation. Only node satisfying interestingness threshold will be included in summary. Following factors are under consideration to calculate summary.

A) proRadius of node in graph.
B) proScope of node into graph

3.2.2 Spreading Tree Model

Although network cascades can model diffusion processes, several issues of dynamics challenge the effectiveness of network cascade model. First, the time-stamped instances of a node are merged in cascades because a node in a cascade model can only appears once. However, in dynamic networks, the interestingness is defined on node instances. Namely, a node becoming interesting is associated with a specific time in an interaction. To distinguish node instances in different interactions and cascades, a cascade model requires a labeling scheme as extra effort. Furthermore, as cascades are directed graphs, there exist backward and forward edges or even cycles. This makes a cascade hard to interpret and navigate between node instances. Second, summary search on cascades can be regarded as subgraph search. However, graph search is usually time consuming since it involves isomorphism checks. Third, since cascades are merged into one directed graph, the graph search space grows exponentially, which makes dynamic summarization even harder. We propose to instead use a Spreading-Tree model. First, spreading trees are constructed directly by interactions without any other efforts. A spreading tree models an individual diffusion process. Information is diffused from the root to the leaves. The model distinguishes interactions and cascades by itself. Next, tree search is relatively efficient. Numerous proposals of efficient tree operations (e.g., update) exist. Third, there are no backward and forward edges in spreading trees. The tree structure is not as complex as a cascade. The search space is proportional to the scale of the interactions. the spreading-tree model achieves the following properties: 1) cascades can be modeled as spreading trees such that the summarization on cascades equals the task on spreading trees; 2) the trees are separated by seeds;3) a node can be duplicated in a spreading tree, which shows that the model distinguishes node instances; 4) the size of the trees is proportional to the scale of the interactions;5) infection occurs top-down, and diffusion occurs from a parent node to a child node.

3.2.3 Self Adjusted parameters:

Although using fixed values for parameters is simple to implement, there are two main issues that demand better approaches. First, for a single diffusion process, prediction of the network statistics (change rate, number of infectees, propagating range, etc.) is usually difficult. It is hard to find parameter settings that can best capture the dynamics. Second, different diffusion processes vary substantially in range and scope. Thus, the same settings are not likely to

work across different processes. Our study aims to provide a self-tuning mechanism that adapts to differences in the summarization of dynamics

1. Alpha Estimation

The minimum a turns out to produce the most informative summaries. The summarization can therefore adapt dynamically.

2. Threshold Selection

Our goal is to find a proper threshold that can make the summarization converge fast and produce a small sized summary over time. However, the changes and differences of dynamics challenge the setting of such a threshold. Therefore, a selection mechanism adapting to the trends of dynamics (i.e., rise and fall) is necessary.

3.2.3.4 Algorithm

Input : Network G ,seed set I(G).

Output : A set of summarized spreading trees, S(G)

1: Begin

2: Threshold $\tau \leftarrow 0; \alpha \leftarrow 0; n \leftarrow |V|$

3: Boolean breakFlag ←false;

4: List path←null;

5: Spreading tree set Set(T) rooted by seeds in I(G).

6: if breakFlag= =false then

7: if x= (δ, V_i, V_j, t) ←$D(G)[t_{ij}]$*exists then*

8: $T \leftarrow$ mapT(Set(T), δ*);*

9: V_i←Search(T, V_i);

10: branchOut(V_i, V_j,t);

11: if $\xi(v_i) > \tau$ then

12: $\tau \leftarrow \xi(v_i)$,*set* α*;*

13: While v_i.*getInfector(T)*ЄI(G) do

14: path.Push(v_i.*getInfector(T)*);

15: S(T) ←getST(Set(T),T);

16: insertPath(S(T),path);

17: **return** Set (T);

Algorithm captures two essential aspects of OSNet: 1)Constructing spreading trees (lines 7 t10); 2) summarizing the most interesting dynamics into S(T) (lines 13 to 16). We proceed to explain the details. We allow users to terminate a summarization process through the variable breakFlag in line 6. Depending on the applications, one can also use a bound on the size of S (T)to abort the algorithm. Note that we have no limitation on n. Once a new interaction x = (δ, u, v, t) arrives (line 7), we call mapT in line 8 to retrieve the T of story d. Next, branchOut in line 10 inserts an infectee v_j from vi with edge label t into T. We implement each S(T) as a search tree. From line 11, we summarize the updated node according to Equation (1). If the node's interestingness exceeds the threshold, it shows a diffusion rise, and the node is inserted into S (T). Parameters are adjusted in line 12.

4. CONCLUSION

All previous network inference algorithms have assumed diffusion networks to be static. Therefore, they have considered the pathways over which information propagates to be static over time. In contrast, we developed a framework for time-varying network inference, OSNet. This provides online time varying estimates of the edges of the network as well as the dynamic edge transmission rates, which allows us to detect how information pathways emerge and vanish over time. There exist limitation and open questions for this study, which however points several promising directions for future work. For instance, 1) dynamic addition of seed nodes and unobservable diffusion processes are beyond the consideration of this study; 2) this work has no assumption on underlying networks that may have influence on the effectiveness of summarization; 3) this study assumes simple cascades as information diffusion, which could be extended to support other models considering user influences and explicit time granularity of interactions; 4) definitions of interestingness may be application-oriented, which may require a generalized version of OSNet.

REFERENCES

[1] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs,"in Proc. SIAM Int. Conf. Data Mining, 2007, pp. 551–556..

[2] F. Zhu, Q. Qu, D. Lo, X. Yan, J. Han, and P. S. Yu, "Mining top-k large structural patterns in a massive network," in Proc. VLDB Endowment, vol. 4, no. 11, pp. 807–818, 2011.

[3] Q. Qu, C. Chen, C. S. Jensen, and A. Skovsgaard, "Space-time aware behavioral topic modeling for microblog posts," IEEE Data Eng. Bulletin, vol. 38, no. 2, pp. 58–67, Jun. 2015..

[4] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the Flickr social network," in Proc. 18th Int. Conf. World Wide Web, 2009, pp. 721–730.

[5] Q. Qu, F. Zhu, X. Yan, J. Han, P. S. Yu, and H. Li, "Efficient topological Olap on information networks," in Proc. 16th Int. Conf. Database Syst. Adv. Appl., 2011, pp. 389–403.

[6] Q.Qu, S. Liu, C. S. Jensen, F. Zhu, andC. Faloutsos, "Interestingness driven diffusion process summarization in dynamic networks," in Proc.Mach. Learn. Knowl.Discovery Databases, 2014, pp. 597–613

[7] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," ACM Trans. Knowl. Discovery Data, vol. 5, no. 4, Art. no. 21, 2012.

[8] M. Gomez-Rodriguez, D. Balduzzi, and B. Sch€olkopf, "Uncovering the temporal dynamics of diffusion networks," in Proc. 28th Int. Conf. Mach. Learn., 2011, pp. 561–568.

[9] M. Gomez-Rodriguez, J. Leskovec, and B. Sch€olkopf, "Structure and dynamics of information pathways in online media," in Proc. 6th ACM Int. Conf. Web Search Data mining, 2013, pp. 23–32.

[10] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," ACM SIGMOD Rec., vol. 42, no. 2, pp. 17–28, 2013

[11] Nargis Pervin, Hideaki Takeda, Fujio Toriumi, , "Factors Affecting Retweetability: An Event-Centric Analysis on Twitter," Thirty Fifth International Conference on Information Systems, Auckland 2014

[12] Chunxiao Jiang_†, Yan Chen_, and K. J. Ray Liu_"Evolutionary Information Diffusion over Social Networks", arXiv:1309.2920v1, Sep 2013

[13] Didier Henrya, Erick Stattnera, Martine Collarda, "Social media, diffusion under influence of parameters", ELSEVIER, The 8th International Conference on Ambient Systems, Networks and Technologies ,2017