Study of Effect of Node Seniority in Social Networks

Baojun Qiu*, Kristinka Ivanova†, John Yen† and Peng Liu†
*Department of Computer Science and Engineering
†College of Information Sciences and Technology
The Pennsylvania State University, University Park, PA 16802
Email: {bqiu, kivanova, pliu, jyen}@ist.psu.edu

***Abstract*—In evolving social networks, nodes join, make connections, or leave over time. In this paper, we introduce node event sequences that record activities of every node over time. Node event sequences are suitable for microscopic analysis of node behaviors in social networks. As preliminary results of taking advantage of node event sequences (as well as snapshots of networks), we study the health of the community of Nanotechnology based on the analysis of $Seniority$ of nodes, identify the intrinsic dynamics of formation of new edges and its relation to $Seniority$ of nodes, and the changes in the node behavior according to the nodes' $Seniority$.**

I. Introduction

In an evolving social network, the set of nodes and the set of edges change over time due to new nodes joining, old nodes leaving, and the formation of new connections. Many studies on evolution of large scale networks work on snapshots and focus on macroscopic properties and their evolution including degree distribution, diameter, centrality, community structure, etc [1], [2], [3], [4]. However, snapshots are not convenient for microscopic analysis [5], such as studying the behavior change of nodes over time. For example, in scientific collaboration networks, researchers usually publishes papers in collaboration with more senior researchers when they are junior, and to help more junior researchers as they gain experience in the field.

In this paper, we introduce NES (node event sequence) that contains all the events involving the node in the order of time, e.g., when a researcher (node) joins a field and when he publishes papers and with whom (more details in Section II). The node event sequence is different from the edge creation sequence [5], in that an event in a node event sequence contains richer information and may correspond to zero or multiple edges. For example, if a researcher publishes a paper with two colleagues, then the event corresponds to 3 edges in the collaborative networks. Also, we can annotate the edges based on information from the paper, e.g., which venue the paper is published and what the topic is. With the node event sequence, it is possible to apply methods developed in temporal / sequence data mining to the microscopic analysis of the evolution of social networks. In this paper, we work on both snapshots and the node event sequences and present some preliminary results to this direction, focusing on study of effect of seniority on network dynamics. Such dynamics could play an important role in various intelligence analysis scenarios [6], [7].

In this paper, we study the nanotechnology collaboration network *NanoSCI*, which is appealing for investigating social network growth dynamics for the following reasons. Firstly, collaboration networks have been widely used in scientometrics and social networks studies. Collaboration networks have been learned to possess many static and dynamic properties that are similar to other social networks[8], [9]. Secondly, *NanoSCI* offers an extensive database including $292,323$ researchers and $368,511$ papers that are indexed by SCI (Science Citation Index) database spanning from 1980 to 2006.

II. Node Event Sequence

We denote an event as $E_k = (\{v_1, v_2, ..., v_x\}, t^k, info)$ where $v_i,\ 1 \leq i \leq x$ represents nodes participating in event E_k, t^k indicates the occurrence time of event E_k and $info$ indicates the description of the event. The set of nodes is denoted as $V(E_k)$ and the time of event is denoted as $T(E_k)$. We use ES(t) to denote the set of events occurred before and at time t, and VS(t) denotes the set of nodes in all events in ES(t). We use $\Delta ES(t_k)$ to represent the set of new events occurred at time t_k. The corresponding set of new nodes is $\Delta VS(t_k)$. We have $ES(t_k) = ES(t_{k-1}) \cup \Delta ES(t_k)$ and $VS(t_k) = VS(t_{k-1}) \cup \Delta VS(t_k)$. The events involving node v at time t is defined as $ES(v,t) = \{E_i | E_i \in \Delta E(t)\ and\ v \in V(E_i)\}$ The node event sequence (NES) of a node v has the following definition

$$NES(v, t_n) = \{ES(v, t_1),\ ES(v, t_2),\ ...,\ ES(v, t_n)\} \quad (1)$$

where $t_1 < t_2 < ... < t_n$. If there are no new events involving v occurred at time t_i , then $ES(v, t_i)$ is an empty set Φ. If we eliminate all empty sets, we get

$$NES'(v, t_n) = \{ES(v, t'_1),\ ES(v, t'_2),\ ...,\ ES(v, t'_s)\} \quad (2)$$

where $t'_1 < t'_2 < ... < t'_s$, $\Phi \subset ES(v, t'_i)$, $1 \leq i \leq s$ and t'_1 is time of node v joining the network.

From NES', we define node $Seniority$ and $Lifetime$ as

$$Seniority(v, t_c) = t_c - t'_1, \quad Lifetime(v, t_x) = t_x - t'_1 \quad (3)$$

where t_c is current time, t_x is leaving time of node v. We define $Intervals$ as

$$Intervals(v) = \{t'_2 - t'_1,\ ...,\ t'_j - t'_{j-1}\} \quad (4)$$

Within collaborative networks, there are often no explicit signs of node leaving. If a node has been inactive for a long enough time η, we may decide that the node has left the network. We set η for $NanoSCI$ in section III-A based on analysis of the $Intervals$.

In the context of collaborative networks, we define an event as researchers publishing a paper. Depending on the number of coauthors, 0 or multiple edges are added in the collaborative network. A weight can be put on the edges according to the total number of the coauthors or the order of the coauthors. Annotations can also be added on the edges based on the paper's content if needed.

Node event sequences provide a very different perspective from what snapshots can support. A snapshot at time t_k is $G(t_k) = \{V(t_k), E(t_k)\}$, which is a static graph [10] where $V(t_k)$ is vertex set and $E(t_k)$ is the edge set. Node event sequences are node oriented and suitable for node behavior study, while snapshots are network oriented and suitable for study of global network properties. In this paper, we use both node event sequences and snapshots to study effect of $Seniority$ in $NanoSCI$.

III. Empirical Study of Node Event Sequences

In this section, we present the results from the empirical study of $NanoSCI$. We extract node event sequence of all nodes and use the sequences, as well as snapshots to understand the dynamics of social networks.

A. Defining Active Nodes

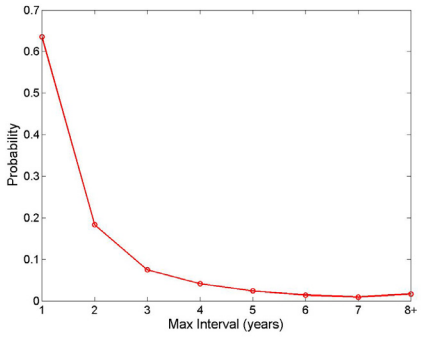

Fig. 1. Distribution of the intervals between two consecutive events on event sequences of researchers (nodes)

In scientific collaborative networks, there is no explicit sign of node leave taking. This is the case for many other networks. How to decide whether a node is active or not? How long a node that exhibits no activities can be regarded as inactive? Using the node event sequences, we can get the distribution of intervals between the nodes' consecutive activities. Since we are interested in finding the threshold of interval for deciding with significant certainty whether a node has left a network, we calculate the set of maximum intervals in the event sequences of all nodes in the snapshot of year 2006 as

$$\{\max(Intervals(v))|\forall v \in V(2006)\} \quad (5)$$

where $Intervals(v)$ is defined in Equation (4) and V(2006) is the set of nodes in the snapshot. In Figure 1, we show the distribution of maximum intervals of all nodes. From the figure, we can see that more than 60% of the researchers have collaborations every year, and about 90% of the researchers have the maximum intervals equal to or smaller than 3 years. Therefore, for this dataset, we decide that a node has left the $NanoSCI$ network if it has been inactive for 3+ years.

B. Seniority and Research Lifetime

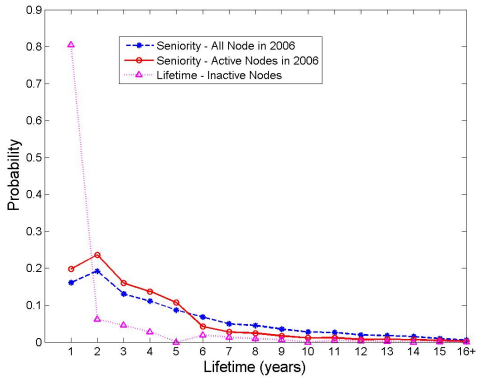

Fig. 2. Distribution of seniority and research lifetime

In this subsection, we study the length of active time of researchers. We have defined $Seniority$ and $Lifetime$ in Equation 3 based on node event sequences. The blue dashed curve in Figure 2 show the $Seniority$ distribution of all nodes in the network snapshot in 2006. However, this may not be accurate because it includes nodes that may have left the network (have been inactive for a long enough time, e.g., 3+ years). The red solid curve shows only the $Seniority$ distribution for active nodes in 2006. This shows that a significant percentage (80+%) of active nodes in 2006 has joined the field as recently as 5 years. Also, the numbers of younger researchers increase against $Seniority$. This suggests that the community of nanotechnology is healthy, and has been still developing rapidly in recent years because newcomers keep joining. Note that the number of nodes added in 2006 is smaller than number of new nodes joined in 2005, this is because our data set was obtained in August, 2006 and does not contain data for the whole year of 2006. The magenta dotted curve shows the lifetime distribution on nodes that may have "left" the network (with 3+ years of disappearance). It shows that about 80% of the researchers have been active in collaboration for only 1 year and then leave the field. We speculate that this is because there are a lot of students in the field. Most of them may graduate after 1 year on publishing and leave the research community.

C. Seniority and Formation of New Collaborations

In this section, we study the relationship between the formation of new collaborations (edges) and the $Seniority$ of researchers (nodes). We obtain the percentages of active

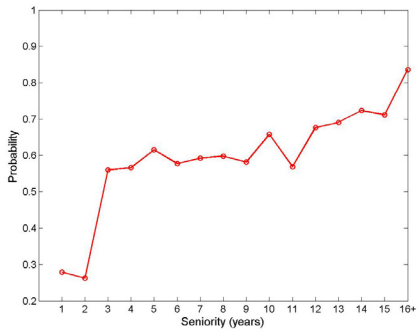

Fig. 3. Formation of new collaboration (edges) as a function of the seniority of the researchers (nodes)

nodes with different $Seniority$ that form new edges in the snapshot of 2006.

Figure 3 shows the results of $Seniority$ against attraction of new connections. It suggests that senior researchers are more likely to get new collaborations while a relative very small percentage of junior researchers ($Seniority \leq 2$) attract new connections in 2006. This could be a supplement for our previous growth model [10].

D. The Interaction Between Senior and Junior Researchers in Publication

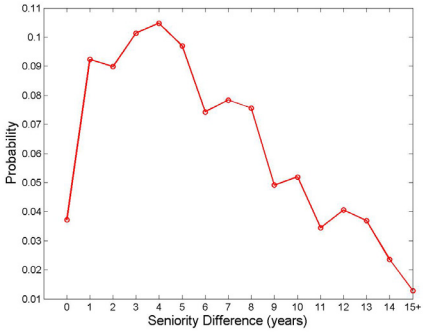

Fig. 4. Distribution of seniority difference between the most senior and the most junior coauthors of a paper

In this subsection, we study the $Seniority$ difference between the most senior and most junior coauthors of every paper published in 2006. Note that less than 5% of the papers are published by only one author and they are not included. Figure 4 shows the results of $Seniority$ difference. We see that nearly 80% of the papers are authored by a junior researcher with at least one more senior researcher of at least 3 more years' experience. The most likely $Seniority$ difference between coauthors is 4 years. These observations suggest that usually senior researchers lead junior ones to do research in the field. They also suggest that researchers may change their roles from getting help to providing help for research as they gain more experience. Notice that the coauthorship with $Seniority$ difference equal to 0, 1 and 2 is not negligible. It also makes sense that some papers are published by authors with similar $Seniority$ and all of whom are experienced.

IV. Conclusion

We define node event sequence (NES) that makes it convenient to carry out microscopic analysis of evolving social networks such as node behavior evolution. In this paper, we mainly focus on studying the effect of node $Seniority$ on network dynamics. More specifically, we have done a case study in $NanoSCI$. We distinguish active nodes and inactive nodes in the network based on the analysis of $Intervals$ in node event sequences; study the health of communities based on the analysis of $Seniority$ of researchers; identify the relationship between formation of new collaborations and $Seniority$ of researchers, and the possible behavior changes of researchers according to their $Seniority$. There are lots of possible extensions to this work, such as evaluating our studies on more datasets and incorporating $Seniority$ to growth models. Also, we can further define other types of sequences from NES according to our purposes. For example, sequences containing numbers of events participated by every node in every time step may be useful for studying activeness of nodes; or sequences containing the types of events involved by an actor may be useful for studying evolving interests of the actor. Furthermore, we can also classify nodes based on their similarity on interest shifts. The work can be naturally extended to other complex network research areas that may relate to homeland security studies, for example, evaluating the anti-social tendencies of key persons.

Acknowledgment

The authors thank Jonathan H. Morgan for useful comments and the funding support from Defense Threat Reduction Agency under contract HDTRA1-09-1-0054.

References

[1] M. Newman, *The structure and function of complex networks*. SIAM Review, 45, 2:167C256, 2003

[2] R. Kumar, J. Novak, and A. Tomkins, *Structure and evolution of online social networks*. SIGKDD, pp. 611-617, 2006

[3] H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, J. Yen, *An LDA-based Community Structure Discovery Approach for Large-Scale Social Networks*. IEEE-ISI, pp. 200-207, 2007.

[4] J. Leskovec, J. M. Kleinberg, and C. Faloutsos, *Graph evolution: Densification and shrinking diameters*. ACM TKDD, 1(1):2, 2007.

[5] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, *Microscopic evolution of social networks*. SIGKDD, pp. 462-470, 2008.

[6] M. Sageman, *Understanding Terror Networks*. University of Pennsylvania Press, Philadelphia, 2004.

[7] Hsinchun Chen, *Homeland security data mining using social network analysis*. IEEE-ISI, 2008

[8] M. E. Newman, *Coauthorship networks and patterns of scientific collaboration*. PNAS, 101 Suppl1:5200–5205, 2004.

[9] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, *Evolution of the social network of scientific collaborations*. Physica A, 311:3, pp. 590-614 2002.

[10] H. Zhang, B. Qiu, K. Ivanova, H. C. Foley, C. L. Giles, J. Yen, *Locality and Attachedness-based Temporal Social Network Growth Dynamics Analysis*. Journal of the American Society for Information Science and Technology, To appear.