Summarization and Visualization Of Communication Patterns in a Large-Scale Social Network

Preetha Appan[1], Hari Sundaram[1] and Belle Tseng[2]

[1] Arts Media and Engineering Program, Arizona State University, Tempe AZ 85281.
{Preetha.Appan, Hari.Sundaram}@asu.edu
[2] NEC Research, Cupertino CA.
belle@sv.nec-labs.com

Abstract: This paper deals with the problem of summarization and visualization of communication patterns in a large scale corporate social network. The solution to the problem can have significant impact in understanding large scale social network dynamics. There are three key aspects to our approach. First we propose a ring based network representation scheme – the insight is that visual displays of *temporal dynamics* of large scale social networks can be accomplished *without using graph based layout mechanisms.* Second, we detect three specific network activity patterns – *periodicity, isolated* and *widespread* patterns at multiple time scales. For each pattern we develop specific visualizations within the overall ring based framework. Finally we develop an activity pattern ranking scheme and a visualization that enables us to summarize key social network activities in a single snapshot. We have validated our approach by using the large Enron corpus – we have excellent activity detection results, and very good preliminary user study results for the visualization.

1. Introduction

This paper deals with the problem of summarization and visualization of large scale social network communication patterns. Understanding large scale social networks is an emerging area of research [9]. The problem is made difficult due to the large size of the network and the long term duration of these networks. Hence visualization and summarization tools that enable users to gain insight into the dynamic behavior of these networks are extremely important.

There has been extensive work in visualization of graph data. Various graph layout algorithms have been developed to enable exploration of large graphs [6]. However these visualizations are for a single large scale graphs. Tools developed to visualize graph data that change over time, show only one graph at a single time instance with a slider to move the graph forward / backward in time. However understanding the temporal dynamics in

the network is difficult. Prior work in analysis of communication has focused on issues such as the information propagation in blogs [5] and community structure detection. However prior work does not explore email communication patterns that are influenced by both time and people. There also has been little focus on summarizing key social network activity patterns though visual means. There has been prior work in innovative visualizations for data analysis [4,7]. We focus on two aspects not addressed before – (a) closely coupling the results of the visualization to the specifics of the social network activity patterns, (b) providing a systematic framework for summarizing the entire social network communication predicated on a topic.

We address the summarization and visualization problems by solving three sub-problems. (a) defining a ring visualization framework for social network activity representation. (b) detecting and visualizing three specific activity patterns and (c) providing a single snapshot summary of the entire network activity. Our visualization framework is inspired

Figure 1: Summarization: Ripples in water provide a compact snapshot view of temporal activity.

by the observation that natural phenomena (ref. Figure 1) can compactly summarize long term activity. The key insight is that compact representation of large scale networks, need not require graph based visualizations. We develop a ring based visualization and summarization framework, that displays relationships between people, time and topic.

We detect three specific activity patterns – *periodicity*, *isolated events* and *widespread growth* at multiple time scales and develop specific ring based visualizations for each activity. The summarization framework allows us to represent key activity patterns over the entire duration of the network in a compact manner. Periodic patterns in time are detected using local maxima of message activity. Regularity in people refers to people who appear frequently in the conversation – this is detected using set intersection techniques. Isolated patterns are detected using constrained global maxima detection, while distributed growth can be detected using a multi-scale message activity analysis (more details in [2]). We have conducted experiments over the large Enron corpus, and preliminary user studies on the visualization, with excellent results.

The rest of this paper is organized as follows. In the next Section we present our approach to visualization of large scale network activity. In Section 3, we discuss our activity pattern detection algorithms. In Section 4 we present out summarization algorithm. In Section 5 we discuss our experimental results and then present our conclusions.

2. The Visualization Problem

In this Section we will present our visualization framework. The central innovation in our approach is that visual displays of temporal dynamics of large scale social networks can be accomplished without using graph based layout mechanisms.

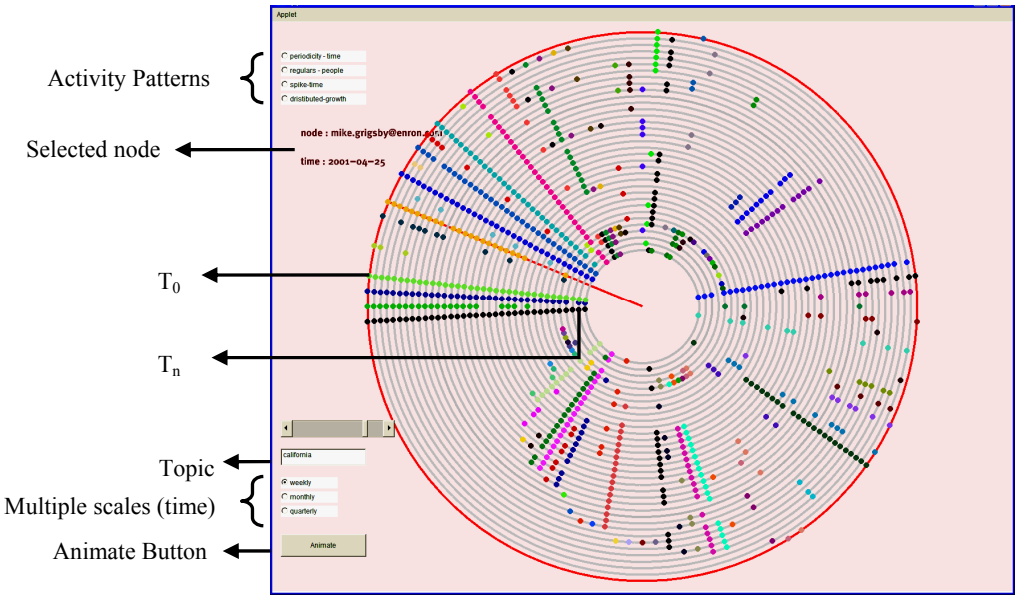

Figure 2: Ring Visualization – Time is represented as concentric circles with the innermost circle representing most recent time slot and outermost ring the oldest time, similar to natural phenomenon. People are indicated with colored dots whose radial location is consistent over time, thus making it very easy to understand how people communicating about a topic vary over time.

2.1 Graphs do not reveal Network Temporal Dynamics

We are addressing the problem of visualizing email communication amongst members of in a large scale social network over an extended period of time. In our system, we are using the Enron Data set. This dataset was collected and prepared by the CALO Project [3]. It contains email data from about 150 users, from senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages. In our system, we create an edge between two nodes (people) if there is evidence of communication between them. Graphs can be very useful to understand the structural

properties of any social network – i.e. who talks to whom.. However, graphs do not reveal the temporal dynamics of the communication in the social network. This is because a graph represents the state of the network at a *single* time instant. This can be a significant problem in large networks such as the Enron corpus that has large user set (150) and significant communication activity over a long duration. Simple techniques such as animation, graph aggregation will not work well.

2.2 Our Approach: Rings

The intuition behind our visualization comes from observing phenomena in the physical world. For example, as seen in Figure 1 we observe ripples in water start from the center and radiate outward. We observe that the growth or spread of energy in these phenomena happens in a radial direction starting from the innermost towards the outermost ring. This creates two constraints – (a) temporal: the outermost ring / ripple represents the earliest temporal event. (b) rotational: the relative orientation of each ring is not arbitrary – a line from the center to the outermost ring corresponds to a direction of energy flow. Before we describe the visualization, we briefly discuss message preprocessing.

Message preprocessing: We assume that the user provides a topic, i.e. a keyword. We then find all messages in the corpus relevant to the given topic. Since our focus is on the visualization rather than information retrieval, we are using a simple keyword match algorithm on the subject line to find all messages pertaining to a given topic. The messages obtained are then ordered in time as well as associated with a set of people - the sender and the set of recipients. In our system, users can browse through three scales of time – 'weekly', 'monthly' and "quarterly". We divide the users into three categories from their email ids – (a) network members – employees amongst the 150 users whose emails contributed to the data set (b) other employees – other employees of Enron not part of the initial 150 people and (c) external – people outside of Enron.

Design Elements: We designed a visualization that indicates multiple graphs that vary over time, in a single snapshot (Figure 2). We now explain the design elements of our visualization.

- *Time*: In our visualization, time is represented as concentric circles, with the innermost circle indicating the latest time slot. Additionally, we can also show activity over multiple time scales.
- *People*: Each person is represented as a distinctly colored dot whose radial location is maintained over different circles. Since people in general form an unordered set, we assign a default ordering along the clockwise direction in the order they *first* appeared in messages sorted in time.
- *Activity:* The message density per time slot is mapped to the color intensity of the circles representing them. Higher the message density, darker the color of the circle that represents the corresponding time slot.

- *Animation:* The visualization can be animated to show the evolution of people talking to each other over time. Time slots indicating more recent activity about the topic are added from the innermost ring and move outward, reminiscent of ripples in water.

The graph structure: The graph structure is not obvious when using rings. We have dealt with this issue by indicating the actual communication graph structure when the user clicks on a particular node in a certain ring. To bring the graph into focus, the rest of the nodes in other time rings are dimmed out by changing their color saturation.

3. Activity Patterns

We now discuss the detection and visualization of three specific temporal communication patterns (periodic, isolated and widespread) in a social network to help summarize the activity with respect to a certain topic. The activity patterns we describe are an extension of the chatter and spiky communication patterns in blogspace that are described in [5]. We add two novel patterns – *distributed growth* and *regulars in people*, to the spiky patterns described there. Also while [5] looked at variations in communication over time, the activity patterns we describe depend on both time and people. We assume that we are given the topic and the corresponding set of relevant messages (ref. Section 2.2).

3.1 Periodic Activity patterns

We shall detect periodic patterns that are regular over time, as well as regular over people. Periodic patterns over time refer to high message activity in the network relevant to a particular topic that appears in regular time intervals.

Detecting periodicity in time: Periodic patterns in time are revealed by detecting the local maxima in message activity and then imposing simple temporal constraints on the maxima. The periodicity detection algorithm proceeds as follows. First all messages are ordered in time, and then grouped according to any chosen scale (weekly, monthly, quarterly). Then each time slot is given an activity score using the following equation:

$$S(t_i) = \sum_{j=1}^{N} P_j(t_i), \qquad (1)$$

Where $S(t_i)$ is the score given to the i^{th} time slot , N is the total number of people involved in messages about the topic, $P_j(t_i)$ is one if the j^{th} node is present in message communication at time instant t_i, zero otherwise. This score is high on time slots involving large number of messages *and* recipients. All the local maxima from the time series scores obtained using eq. (1) are marked as "peaks." Temporal distances are

computed from each peak to every other peak and stored in a table per peak. The local distance tables are then combined to construct a global histogram of distances. We consider a period to be valid only if there exists at least one set of three peaks at the same temporal distance from each other. For example if $d(p_1,p_2) = d(p_2,p_3) = d(p_3,p_4) = d_1$, then d_1 is considered a valid period. This removes spurious maxima. The algorithm gives a list of periods and the number of peaks that are participants with that period.

Temporal periodic patterns are easily understood using rings. Every time period that corresponds to a peak and part of the top three detected period sets is colored with a distinct color.

Detecting regulars in people: Regularity in people refers to the set of people who occur together, frequently, over the duration of the topic. This can be detected using a set intersection algorithm. Consider N to be the total number of people exchanging emails about the topic. We iteratively find all subsets S_k from these N people, that occurred together more than q_1 times over all time slots. The threshold q_1 is fixed according to the time scale. These subsets of people form groups that are the 'regulars' to the topic. To visualize the set people who appear together, all nodes (representing people) in the visualization that are part of the same set are colored with the same color. This is done only after the user explicitly selects this pattern to be revealed. Their radius is also increased and the background is dimmed out in order to prominently display the regulars in the topic.

3.2 Isolated patterns

We now show how we can detect and visualize isolated patterns (also referred to as spikes) over time and people. Isolated patterns over time refer to significant message activity over a short time window. Isolated patterns over people refer to information generators – a small set of people, who contribute to most of the messages.

Detecting spikes in time: A spike in time is characterized by three conditions: (a) there exists local maxima in activity, (b) the message activity exceeds a certain threshold and (c) the activity exhibits a sharp rise and fall in small time duration.

In order to find such spikes in time, we first begin with the ordered set of all messages relevant to the particular topic. We use equation (1) to calculate the score of each time slot, which depends on both the number of messages and the number of recipients per message. The global maxima verified with the above constraints are then visualized. In order to indicate spikes in time for the given topic to the user, we highlight the time slot in which the maxima occurred. We additionally increase the radius of all nodes representing people communicating in that time slot.

Spikes in people: Spikes in people refer to the information generators – a small set of people who send a large percentage of messages relevant to the topic. We now define two measures α (sender coverage) and β (message coverage), to be as follows.

$$\alpha(N_s) = \frac{N_s}{N_0}; \beta(N_s) = \frac{1}{M}\sum_{j=1}^{N_s} m_j, \quad (2)$$

where N_s is the number of unique senders, and m_j is the number of messages contributed by the j^{th} sender. Given a certain threshold for β, we can find a corresponding N_s – the minimum number of senders required to generate those messages. The information generator set is determined by determining N_s using equation (2) such that these values of $\beta \geq \beta_0$ and $\alpha \leq \alpha_0$ are satisfied. These N_s senders are then the *information generators* for the given topic. We determined the thresholds ($\beta_0 = 0.65$ and $\alpha_0 = 0.15$) using a training set [2]. Spikes in people in are indicated our visualization, by increasing the size of nodes that are spikes in people in *all* times that they occur. We also place them along equidistant radial lines. Details of our algorithm to detect and visualize distributed growth can be found in [2].

4. Summarization

In this Section we discuss the problem of summarizing the key activity patterns in a single snapshot. The solution involves two steps – (a) the detection and ranking of activity patterns and (b) developing a single representative snapshot. The visualization problem is difficult since the activity patterns need not co-occur within the same time window.

4.1 Ranking the activity patterns

Each activity detector (ref. Section 3) returns a set of detected activities. We thus need to develop measures to order the activities within each set. We now discuss a systematic ranking measure for each activity pattern.

- *Periodic in time*: each period is associated with a frequency (the number of message activity peaks that are in that period), which is used to pick the top three periods.
- *Regulars in people*: Each set of people has a corresponding frequency which is the number of times they occurred together. We use the *average* closeness centrality [8], as a measure to rank the different 'regular sets' of people.
- *Isolated in time*: The message activity score from equation (1) implicitly ranks the sets of the spike in time patterns.
- *Isolated in people*: The *number* of information generators is used to rank the sets. Smaller the set of information generators, higher the rank.
- *Distributed growth*: The size of the time window of growth is used as a ranking mechanism. Larger the time window, higher the rank.

We will now discuss how we construct a summary snapshot to indicate all the key activity patterns, given a certain topic.

4.2 Constructing the summary snapshot

Each of the activity patterns detected could occur at different times as well as involve different people. Hence indicating all of them within the same screen is difficult, especially if the time range is bigger than the maximum that can be shown in the available display area. Instead, we have constructed a *representative* summary snapshot that only visually indicates the key patterns for messages of the given topic, but does not correspond to the actual time of when the pattern occurred. This is an interactive

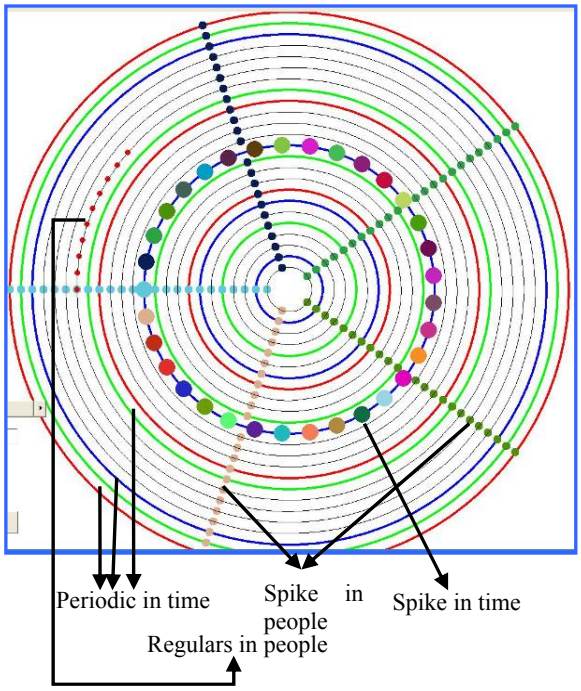

Figure 3: Summary Snapshot – Indicating four different activity patterns in a representative snapshot for the query 'California' grouped monthly.

summary, where the user can then click on the pattern of interest to go to the ring visualization and see the actual time period corresponding to the activity pattern of

interest. The Figure 3 shows an example summary snapshot of the query 'California' in the monthly scale.

5. Experiments

The visualization and summarization framework was implemented in Java and Processing [1] with the Mysql database as the backend. In order to evaluate the system, we conducted a preliminary user study with five graduate students. Users were asked to interact with the system by executing several text queries (e.g. 'power crisis', 'California', 'trading' etc.). They were then asked to answer questions on various aspects of the system on a scale of one to seven. The results are summarized in Table 1 and indicate that users find the activity patterns as well as the visualization to be very useful in understanding email communication. Users also suggested various improvements such as (a) visualizing relationships between a single person, time and the topic, rather than the entire social network, (b) comparing communication activity for multiple topics in the same visualization.

Table 1: Preliminary user studies

Interface Aspects	Score
User Friendliness	**5.0** / 7
Activity Patterns	**6.25** / 7
Multi-scale analysis	**5.75** /7
Helps understand relationships between topics, people and time	**6.0** / 7
Summary snapshot	**6.25** /7

We acknowledge that this is only a preliminary evaluation – the actual study would involve applying our visualization technique to emails from an organization and allowing *members of the same organization* to evaluate whether the visualization was able to communicate temporal patterns well. We also ran activity pattern detection algorithms on 100 queries on the Enron data. The detailed results can be found in [2].

6. Conclusion

In this paper, we proposed a framework for visualization and summarization of email communication activity in large social networks. The framework addressed three challenges (a) visualization (b) activity pattern detection and (c) summarization. The novel ring visualization scheme depicts multiple graphs in the same snapshot and enables users to understand communication activity that varies over multiple scales in time. We also defined and detected three classes of communication activity patterns that depend on people and time – (a) periodic (b) isolated and (c) distributed. We discussed visualization of these patterns using the ring visualization. The detected activity patterns are then summarized by ranking activity patterns and constructing a single snapshot that communicates all key activity patterns to the user. Preliminary experiments and user study results are promising and we plan to conduct further extensive evaluation.

7. References

[1] *Processing* http://proce55ing.net.

[2] P. APPAN, H. SUNDARAM and B. TSENG (2005). *Summarization and Visualization of Communication Patterns in a Large Social network*. Arts Media and Engineering Program, ASU, AME-TR-2005-12, May 2005.

[3] CALO *http://www.ai.sri.com/project/CALO*

[4] J. V. CARLIS and J. A. KONSTAN (1998). *Interactive visualization of serial periodic data*, Proc. of the 11th annual ACM symposium on UIST, 29-38, San Francisco.

[5] D. GRUHL, R. GUHA, D. LIBEN-NOWELL and A. TOMKINS (2004). *Information Diffusion through Blogspace*, Proceedings of the 13th international conference on World Wide Web,

[6] I. HERMAN, M. DELEST and G. MELANCON (2000). *Graph visualization and navigation in information visualization: A survey.* IEEE Transactions on Visualization and Computer Graphics **6(1)**.

[7] D. A. KEIM, J. SCHNEIDEWIND and M. SIPS (2004). *CircleView: a new approach for visualizing time-related multidimensional data sets*, Proc. Advanced visual interfaces, 179-182, Gallipoli, Italy.

[8] M. E. J. NEWMAN (2003). *A measure of betweenness centrality based on random walks.* Social Networks.

[9] J. R. TYLER, D. M. WILKINSON and B. A. HUBERMAN (2003). *Email as spectroscopy: automated discovery of community structure within organizations.* Communities and technologies: 81 - 96.