Localizing Temporal Anomalies in Large Evolving Graphs

Teng Wang *, Chunsheng Victor Fang†, Derek Lin‡, S. Felix Wu*
* Department of Computer Science, University of California, Davis. † Pivotal Software Inc., now at Awake Networks. ‡ Pivotal Software Inc., now at Caspida.

Problem Statement

Temporal anomalies are defined as graph-based outliers presented in one or more instances of dynamic graphs[1, 2]. Localizing temporal anomalies provides the basis to address several application-specific tasks such as insider threat detection in enterprise networks and fraud detection in financial services. However, it is not an easy task. In general, there are two main challenges in this area.

- Event detection vs. localization. Most work focus on event detection[3] and do not address the problem of localization, i.e. identifying the specific nodes that are responsible for the detected anomalous changes in graph structure.
- High false alarm rate. Many non-anomalous nodes may be considered as anomaly due to normal community-level changes.

General Framework

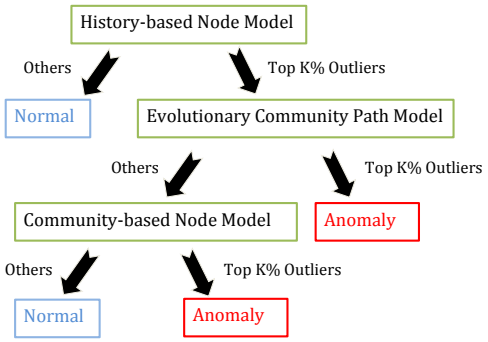

Figure 1: Our Framework for Anomaly Localization.

- History-based Node Model: a node-centric model.
- Evolutionary Community Path Model: a community-centric model.
- Community-based Node Model: a community-centric model.

Contribution

- A novel method based on a Vector Autoregression (VAR) model to localize temporal anomalies in dynamic graphs.
- A new anomaly localization framework with reduced false alarm rate by leveraging both a node-centric model and a community-centric model.
- Performance evaluation on synthetic and real-world datasets covering different application areas including Enron email network data (personal communication), an enterprise network traffic data, and CNN public Facebook page (social media).

Method

History-based Node Model

Feature Matrix. We represent each graph snapshot G_t as a feature matrix:

$$F_t = \begin{bmatrix} g_{11,t} & g_{12,t} & \dots & g_{1n,t} \\ g_{21,t} & g_{22,t} & \dots & g_{2n,t} \\ & \dots & & \\ g_{m1,t} & g_{m2,t} & \dots & g_{mn,t} \end{bmatrix}$$

$g_{ij,t}$ denotes the ith feature value of the jth node at time t. m is the number of selected features, and n is the number of nodes in this snapshot.

Node Behavior Model. We use the Vector Autoregression (VAR) model [4] in time series analysis.

$$F_t^{(i)} = c + A_1 F_{t-1}^{(i)} + A_2 F_{t-2}^{(i)} + \dots + A_j F_{t-j}^{(i)} + \dots + A_p F_{t-p}^{(i)} + e_t \quad (1)$$

c is a $m * 1$ vector of constants and e_t is a $m * 1$ vector of error terms. p is the lag order that can be determined automatically by the VAR model. A_j is a $m * m$ matrix describing the node feature transition between time t and $t - j$:

$$A_j = \begin{bmatrix} a_{1,1}^j & a_{1,2}^j & \dots & a_{1,m}^j \\ a_{2,1}^j & a_{2,2}^j & \dots & a_{2,m}^p \\ \dots & & & \\ a_{m,1}^j & a_{m,2}^j & \dots & a_{m,m}^j \end{bmatrix}$$

Then we define the outlier score of each node in the history-based node model as the difference between the predicted feature vector and observed feature vector. At last, we get *History-based Outlier Score* as $O_{hb,(t+1)}^x = ||\hat{F}_{t+1}^{(x)} - F_{t+1}^{(x)}||$.

Evolutionary Community Path Model

We propose a new definition to evaluate community similarity in all dynamic events, such as forming, dissolving, expanding, contracting, splitting, and merging [5].

$$sim(C_{t,a}, C_{(t-1),x}) = \frac{|C_{t,a} \cap C_{(t-1),x}|}{min(|C_{t,a}|, |C_{(t-1),x}|)} \quad (2)$$

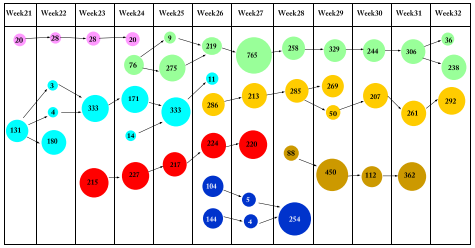

Figure 2: Evolutionary Community Paths in Enron Email Dataset

Similar with Equation (1), we use the VAR model to learn how the activities in community path y changes over time and calculate the *Community Path Ourlier Score* as $O_{cp,(t+1)}^y = ||\hat{F}_{t+1}^{(y)} - F_{t+1}^{(y)}||$.

Community-based Node Model

We define the outlier score of each node in the community-based node model as the difference between its feature vector and the average feature vector of its community. At last, we get the *Community-based Outlier* as $O_{cb,t}^z = ||F_t^{(r)} - F_t^{(z)}||$.

Experiments

We run our framework on both synthetic data set and real world data sets. Some proven real world examples are also used to evaluate the performance of our framework in reducing false alarm rates.

Synthetic Data. We create a graph sequence $G = \{G_t, t = 1, 2, ..., 25\}$ based on an original random graph. In addition, we inject three types of patterns at time $t = 20$ and repeat each pattern injection five times.

The area under the ROC curves for INF, HIB and SCM are respectively given by 0.95, 0.87 and 0.93.

- INF: Our integrated framework
- HIB: History-based node model
- SCM: Model with a static-community algorithm

We find that our integrated framework (INF) performs the best among these three algorithms.

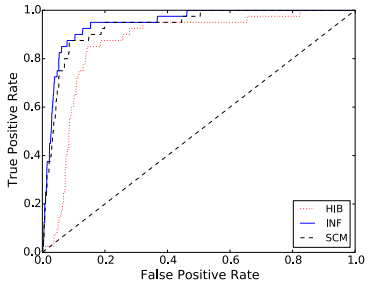

Figure 3: ROC curves comparing different models

Enron Email Network Data. The Enron email network data is a graph structured data set based on emails exchanged among employees of the Enron Corporation from 1998 to 2002. We show some detailed graph feature analysis in Figure 4 and Figure 5. The y-axis of these figures represents the graph feature distribution at each point. Each distinct color represents one graph feature (Black: degree, Blue: clustering coefficient, Green: betweenness, Yellow: closeness, Magenta: eigenvector, Red: PageRank) and the length represents its feature value, which is normalized by the sum of the total features.

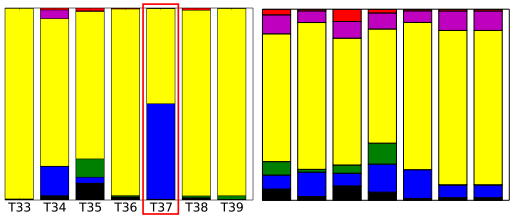

Figure 4: Detected Anomaly. Rosalie Fleming from $T33$ to $T39$ (Left), Community members of Rosalie Fleming at $T37$ (Right). We localize Rosalie Fleming, the assistant to Kenneth Lay, as an anomaly during the time Kenneth Lay returned as chief executive between Aug. 2001 and Sept. 2001.We can easily see anomalous graph feature changes of Rosalie in week 37 (Sept. 17)

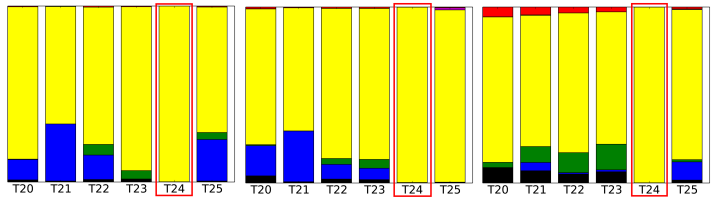

Figure 5: Reduced False Alarms in Enron Email. Danny McCarty (Left), Drew Fossum (Mid) and Stanley Horton (Right). Danny McCarty, the Chief Commercial Officer of Enron's Pipeline Group, shows an unusual inactive email communication pattern at the end of June. 2001. However, by analyzing his dynamic community information, we find that his previous collaborators, Stanley Horton and Drew Fossum, show a similar behavior pattern at that time.

Enterprise Network Traffic Data The enterprise network traffic data consists of user-to-server access traffic records from Jan. 1, 2014 to Jul. 13, 2014.

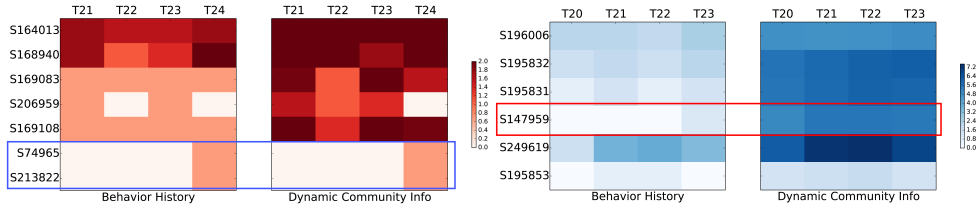

Figure 6: Example Users in the Enterprise Network. Detected Anomaly: User V_1 (Left). Reduced False Alarms: User V_2 (Right). In both of V_1's behavior and its community information, we can find unusual new connections at time $T24$. In V_2's behavior, its unusual connection increase at time $T23$ turns out to be normal in the community information.

References

[1] Rossi, Ryan A and Gallagher, Brian and Neville, Jennifer and Henderson, Keith. Modeling dynamic behavior in large evolving graphs. In *WSDM*, 2013
[2] Sricharan, Kumar and Das, Kamalika. Localizing anomalous changes in time-evolving graphs. In *SIGMOD*, 2014
[3] Akoglu, Leman and Faloutsos, Christos. Event detection in time series of mobile communication graphs. In *Army Science Conference*, 2010
[4] Johansen, Soren. Likelihood-based inference in cointegrated vector autoregressive models. In *OUP Catalogue*, 1995
[5] Greene, Derek and Doyle, Dónal and Cunningham, Pádraig. Tracking the evolution of communities in dynamic social networks. In *ASONAM*, 2010