SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

Monitoring Dynamic Social Networks:

Using SAS/IML®, SAS/QC®, and R

USERS PROGRAM

THE UNIVERSITY OF ALABAMA | Culverhouse College of Commerce

Monitoring Dynamic Social Networks:
Using SAS/IML®, SAS/QC®, and R

Huan Li **Dr. Michael D. Porter**

The University of Alabama , Tuscaloosa AL

ABSTRACT

Dynamic social networks can be used to monitor the constantly changing nature of interactions and relationships between people and groups. The size and complexity of modern dynamic networks can make this task extremely challenging. Using the combination of SAS/IML®, SAS/QC®, and R, we propose a fast approach to monitor dynamic social networks. A discrepancy score at edge level was developed to measure the unusualness of the observed social network. Then, multivariate and univariate change-point detection methods were applied on the aggregated discrepancy score to identify the edges and vertices that have experienced changes. Stochastic block model (SBM) networks were simulated to demonstrate this method using SAS/IML and R. PROC SHEWHART and PROC CUSUM in SAS/QC were applied on the aggregated discrepancy score to monitor the dynamic social network. The combination of SAS/IML, SAS/QC, and R make it an ideal tool to monitor dynamic social networks.

INTRODUCTION

Network data can be challengingly large sometime. And learning the structure of such large network is computationally demanding and it could be even more challenging to dynamically monitor the network. In another word, the high volumes of traffic flowing through these networks could make analysis of network extremely hard. Besides the computational difficulties, large networks also complicate analysis as visualization and inference becomes increasingly impractical.

We are interested in the problem of discovering anomalies in large dynamic networks. Anomaly detection, finding objects, relationships, or points in time that are unlike the rest, is an important problem with variety of applications in different fields, like social science, heath care and cyber security. A key aspect of the networks we consider is that the anomalies only occur in small regions of the network and for particular tine periods. For large networks, global change detection methods can miss the important changes that occur at the subgraph level.

The size and complexity of modern dynamic social networks can mask important changes that occur at the local level. To discover such anomalies, we develop discrepancy measures (e.g., residuals) for all possible edges across time. There discrepancy measures are aggregated to nodes level to identify nodes that have unusual activity.

We then combine these discrepancy measures to find sub-regions, and corresponding time periods, that experienced anomalous activity. We examine several discrepancy measures that can reveal different types of anomalies. SAS/IML Studio is used to access R and simulate data. PROC SHEWHART and PROC CUSUM procedures in SAS/QC were applied on aggregated discrepancy score to monitor the dynamic social network. The combination of SAS/IML, R and SAS/QC makes it an ideal tool to monitor dynamic social networks.

The “Social Graph” behind Facebook

https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919/

THE UNIVERSITY OF ALABAMA | Culverhouse College of Commerce

Monitoring Dynamic Social Networks:

Using SAS/IML®, SAS/QC®, and R

Huan Li **Dr. Michael D. Porter**

The University of Alabama , Tuscaloosa AL

METHOD

i. Edge matrix= $\{e_{ij}(t)\}$ $i, j \in all\ nodes.\ t \in \{1,2,\ldots,T\}$

ii. Index to node

iii. Make degree for each node at every time.

iv. $\widehat{P_{ab}} = \frac{1}{T}\sum e_{ab}(t)$ $\gamma_{ab}(t) = e_{ab}(t) - \widehat{P_{ab}}$

v. $R_a(t) = \sum_{j\ for\ all\ nodes} \gamma_{aj}(t)$

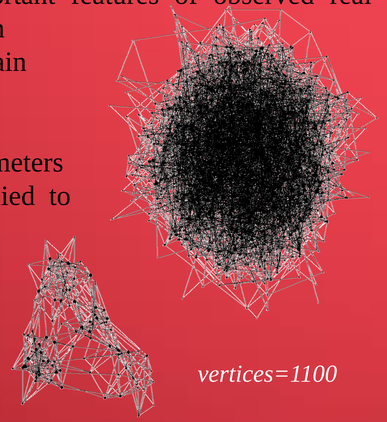
vi. Detect change point of $R_i(t)$ $i = nodes$

vii. Discrepancy/Anomaly visualizing by graphs

SIMULATED DATA

Degree corrected stochastic blockmodel (DCSBM) is proposed as a tool for generating synthetic networks for benchmarks, since it models two important features of observed real networks: 1, community structure and 2, degree heterogeneity, with guaranteed assortative property, dis-assortative property and contain community structure.

A dynamic version of DCSBM is simulated with controlled parameters to mimic the network that undergoes a change. The method is applied to This simulated DCSBM.



vertices=1100

ENRON DATA

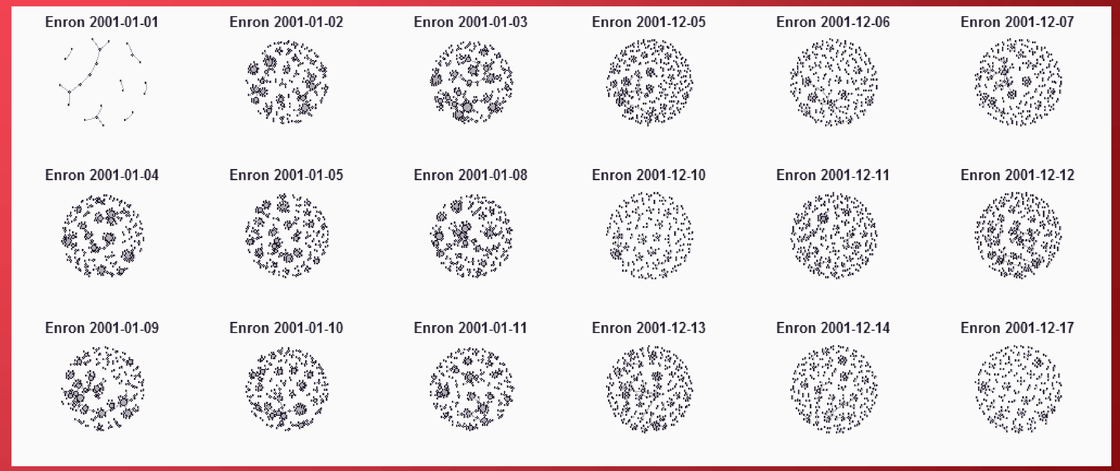
• Enron Corporation was an energy, commodities and services company. Right before Enron's bankruptcy on December 2, 2001, Enron had employed almost 20,000 staff and it was one of the world's major electricity companies.

• In 2001, after disclosures of accounting practices bordering on fraud, the company was forced to file for bankruptcy.

• Enron share prince decreased from $90 USD during the summer of 2000, to just pennies.

• Enron data was originally released by William Cohen at CMU. Different formats of dataset are available to public.

- 2001/01/01 – 2002/04/01
- 22477 vertices
- 456 days/time points ,
- 154 vertices are labelled "core" .



THE UNIVERSITY OF ALABAMA® | Culverhouse College of Commerce

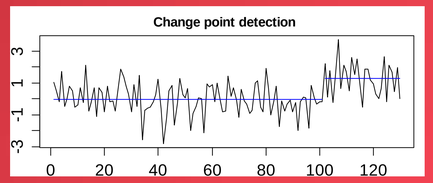
Monitoring Dynamic Social Networks:
Using SAS/IML®, SAS/QC®, and R

Huan Li **Dr. Michael D. Porter**

The University of Alabama , Tuscaloosa AL

SAS/IML

```
submit /R;
x=rnorm(100,0,1);y=rnorm(30,1,1)
a=c(x,y);ansmean=cpt.mean(a)
plot(ansmean, cpt.col='blue')
endsubmit;
```

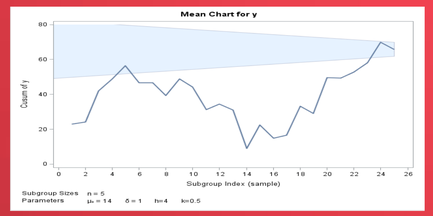


R is a open source statistical software with many add-on packages, providing researchers with very convenient and fast access to the latest research. SAS/IML provides the capability to interface with R.

This simple code showed that SUBMIT/R statement can be used in SAS/IML. And the control chart using Binary Segmentation method (cpt.mean) in R package named “changepoint”.

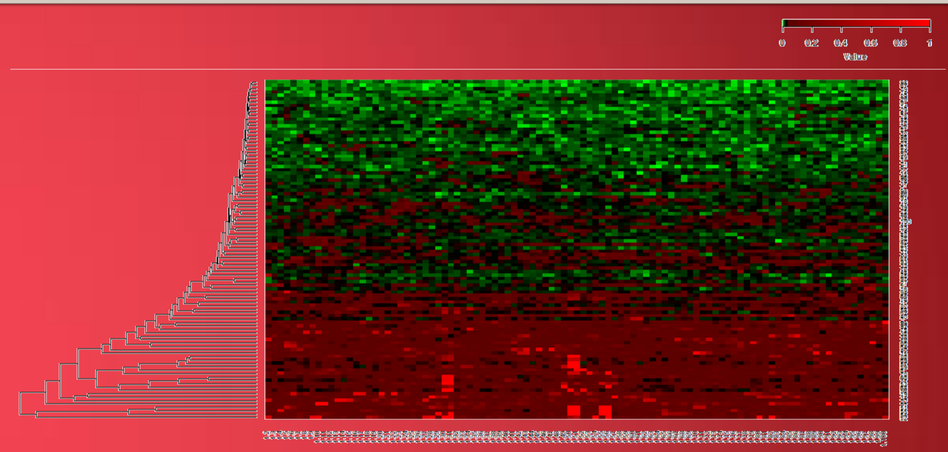
SAS/QC

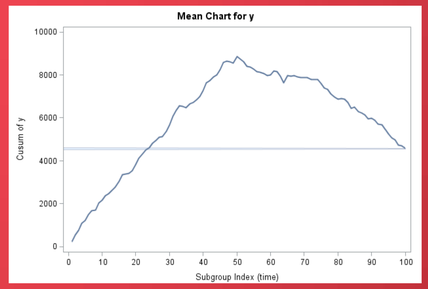
```
title 'c Chart for Change Point';
PROC SHEWHART data=Node_Matrix;
cchart  Score*Node;  run;

title 'Cusum for Change Point';
PROC CUSUM DATA=Node_Matrix;
   XCHART  Score*Node
 / muo=74 sigmao=.005  h=4.0 k=0.5  delta=1.0;run;
```

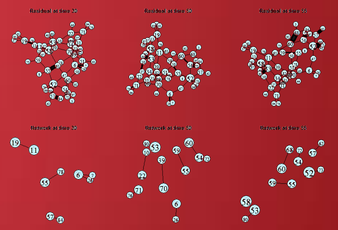


SAS/QC CUSUM procedure creates the Cumulative Sum Control Chart that is widely used for monitoring change detection. SHEWHART procedures creates the Shewhart-control chart.

SIMULATED RESULT







A change is simulated at time 50 impacting 10 nodes out of 1000 nodes. And using the aggregating residual method, we can tell there is a peak at time 50 from the CUSUM chart. Also, heat map of the residual also indicates where and when there are red flags.

Sub-networks at selected time are also showed, compared to the expected sub-networks(with selected nodes) .

THE UNIVERSITY OF ALABAMA® | Culverhouse College of Commerce

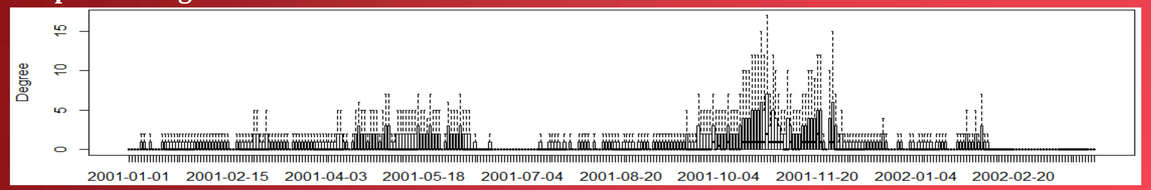
Monitoring Dynamic Social Networks:
Using SAS/IML®, SAS/QC®, and R

Huan Li **Dr. Michael D. Porter**

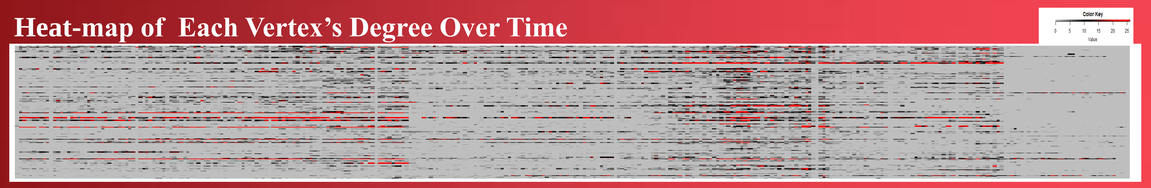
The University of Alabama , Tuscaloosa AL

ENRON RESULT

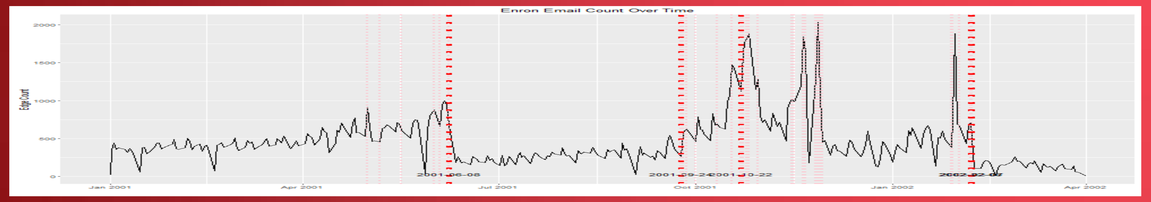
Boxplot of Degree Over Time



Heat-map of Each Vertex's Degree Over Time



Time 1:326



After applying this method, we identified several key moments for the Enron company.
Jun. 6, 2001 – Unknown.
Oct. 22, 2001- Enron acknowledges SEC inquiry into a possible conflict of interest related to the company's dealings with Fastow's partnerships.
Feb 07,2002 – Former chairman and CEO resigned from Enron.

FUTURE WORK

- Estimate edges with weights
  - Hurdle Model
  - Zero Inflated Poisson / Negative Binomial
- Relax independent assumption on edges
  - Time Series Count Model(Autoregressive Conditional Poisson Model, Markov Chain ,)
- Multivariate control chart

REFERENCES

Chandola, V., Banerjee, A. and Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), p.15.
Heinen, A., 2003. Modelling time series count data: an autoregressive conditional Poisson model. Available at SSRN 1117187.
Karrer, B. and M. E. Newman 2011. Stochastic blockmodels and community structure in networks. Physical Review E 83 (1), 016107.
Killick, R. and Eckley, I., 2014. changepoint: An R package for changepoint analysis. Journal of Statistical Software, 58(3), pp.1-19.
Kolaczyk, E.D. and Csárdi, G., 2014. Statistical analysis of network data with R (pp. 1-5). New York, NY: Springer.
Lü, L. and Zhou, T., 2011. Link prediction in complex networks: A survey.Physica A: Statistical Mechanics and its Applications, 390(6), pp.1150-1170.
Newman, M., 2010. Networks: an introduction. OUP Oxford.
Peel, L. and Clauset, A., 2014. Detecting change points in the large-scale structure of evolving networks. arXiv preprint arXiv:1403.0989.
Ranshous, S., Shen, S., Koutra, D., Harenberg, S., Faloutsos, C. and Samatova, N.F., 2015. Anomaly detection in dynamic networks: a survey.Wiley Interdisciplinary Reviews: Computational Statistics, 7(3), pp.223-247.
SAS/QC® 14.1 User's Guide The SHEWHART Procedure
SAS/QC® 13.2 User's Guide The CUSUM Procedure
SAS/IML Studio 14.2: User's Guide

SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL