DynamicDeepFlow: Clustering of Network Traffic Flow Changes using Unsupervised Learning

Sheng Shen Mariam Kiran Bashir Mohammed

Lawrence Berkeley National Laboratory

11/15/2021

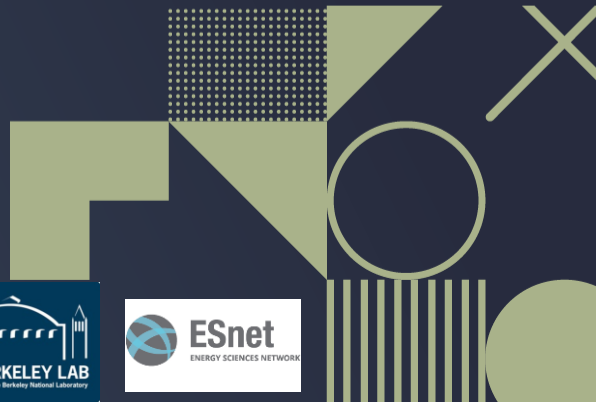
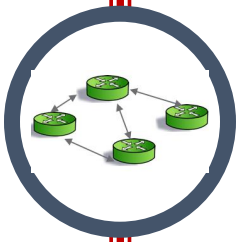


Table of Contents



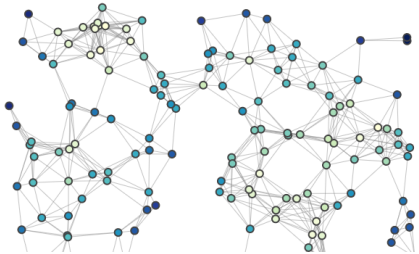
Part 1: Background and Motivation

Part 2: Methodology

Part 3: Experimental Results

Background

Unknown, complex data-to-traffic pattern relationship for network



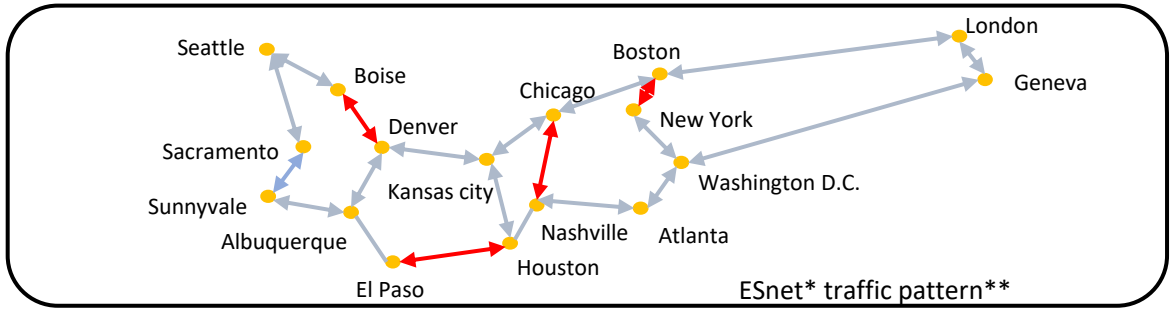
Urgent demands for accurate real-time network traffic pattern monitoring.

Tens of thousands of measurement data points from offline testing and onboard sensing

Motivation

Can we online recognize anomalous network traffic patterns by leveraging large amount of data?

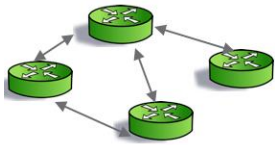
Motivation



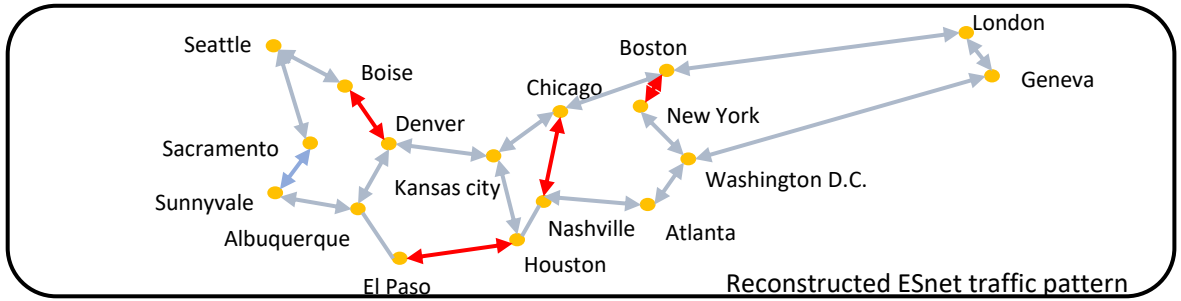
Clear network traffic

Congested network traffic

Deep Learning



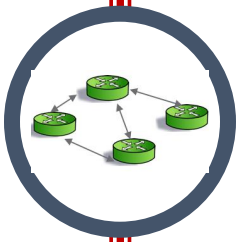
Unknown data-to-network-pattern relationship



* ESnet: Energy Sciences Network

** Some network sites are not marked for a better showing.

Table of Contents



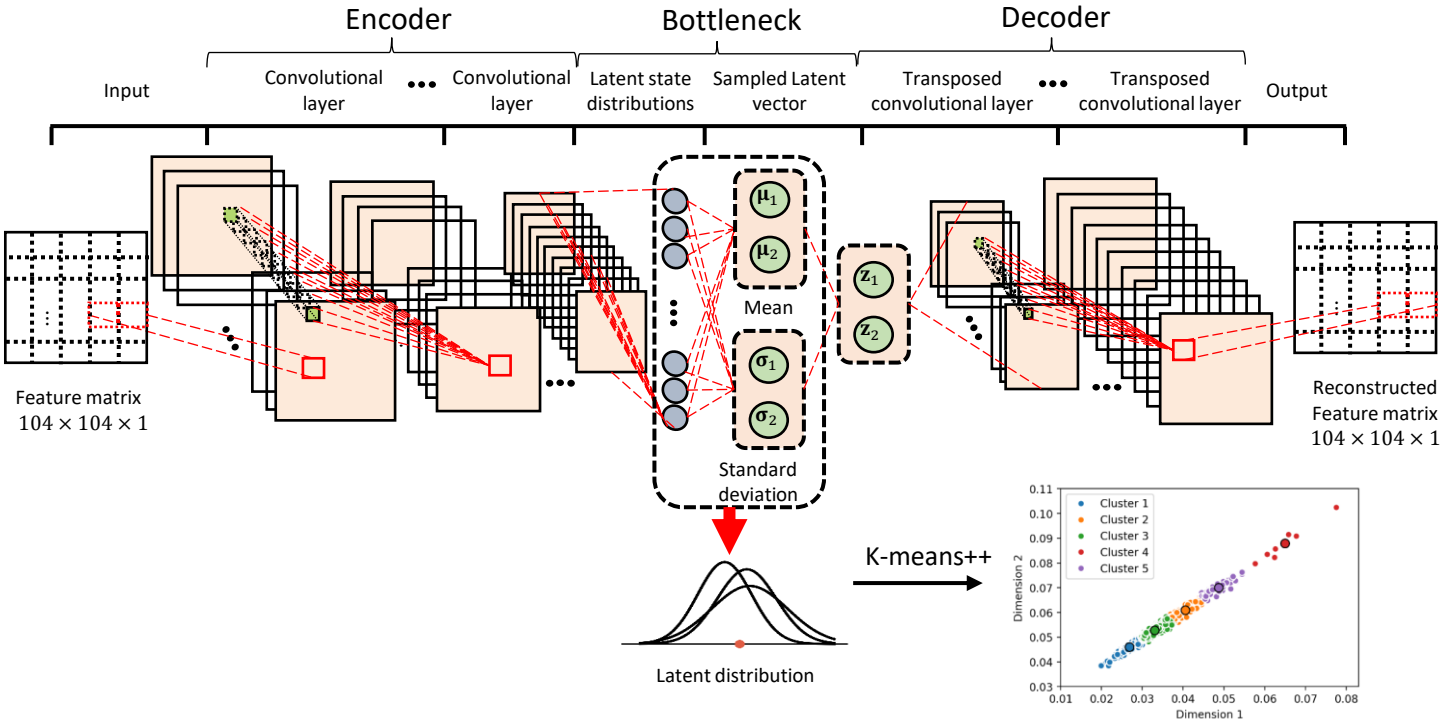
Part 1: Background and Motivation

Part 2: Methodology

- *Developed a machine learning-based network traffic pattern clustering method that incorporates a shallow learning model and a deep learning model.*

Part 3: Experimental Results

DynamicDeepFlow (DDF) model structure



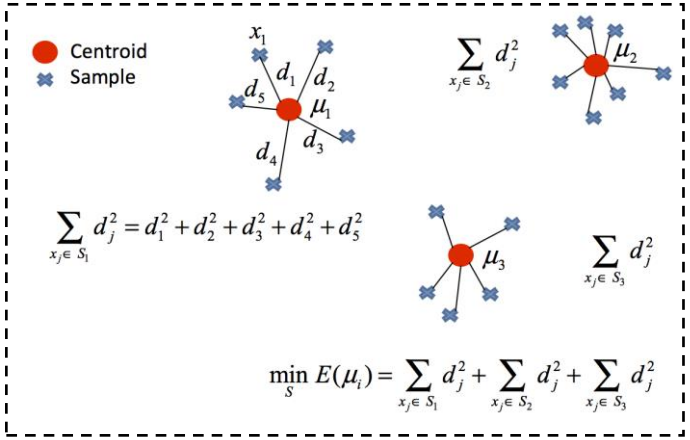
- The DDF consists of a deep learning variational autoencoder model and a shallow learning k-means++.
- The autoencoder automates the feature extraction and avoids risk of dropping useful information in the data using manual feature extraction.
- The K-means++ determines k clusters for normal and anomalous network traffic patterns.

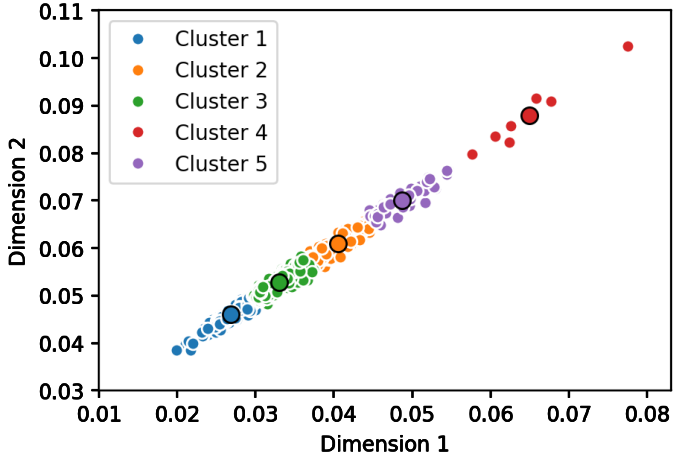
Architecture of variational autoencoder (VAE)

Layer name	Filter size	Number of kernel	Sride size	Number of weights	Number of biases
Input	1×104×104	-	-	-	-
Conv-1	1×5×5	32	(3,3)	800	32
Conv-2	1×4×4	64	(2,2)	32,768	64
Conv-3	1×4×4	128	(2,2)	131,072	128
Conv-4	1×3×3	128	(1,1)	147,456	128
Conv-5	1×3×3	128	(1,1)	147,456	128
FC-1	2×1	-	-	2304	0
FC-2	2×1	-	-	2304	0
FC-3	1152×1	-	-	2304	0
Transposed conv-1	1×3×3	128	(1,1)	147,456	128
Transposed conv-2	1×3×3	128	(1,1)	147,456	128
Transposed conv-3	1×4×4	64	(2,2)	131,072	64
Transposed conv-4	1×4×4	32	(2,2)	32,768	32
Transposed conv-5	1×5×5	1	(3,3)	800	1
Output	1×104×104	-	-	-	-

The overall architecture of the VAE consists of five convolutional layers, five transposed convolutional layers, and three fully-connected layers.

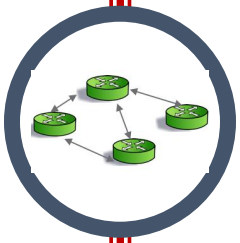
K-means++





- The clustering centers are determined by iteratively minimizing the average squared distance between data points and the clustering center.
- The k-means++ is used to explore the structure hidden in the features and determine *k* clusters for normal and anomalous network traffic patterns.

Table of Contents

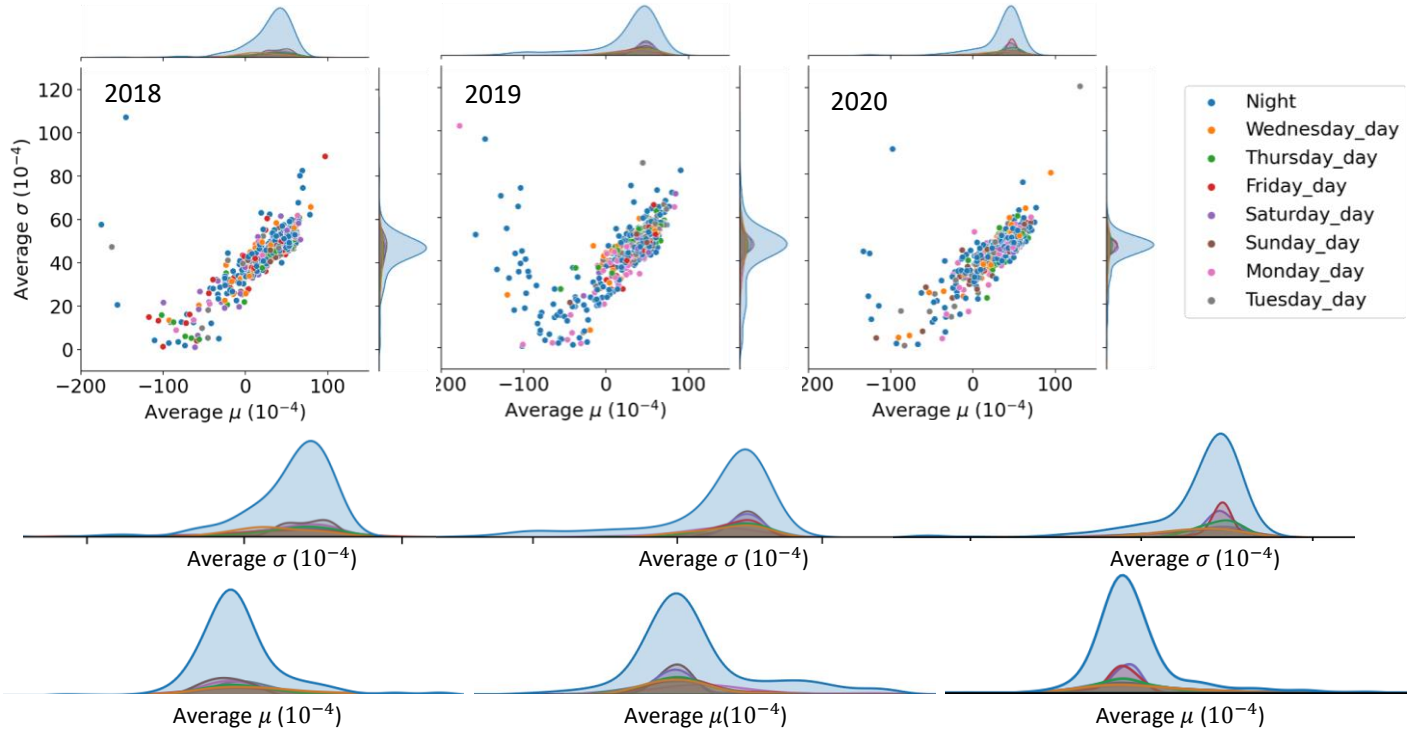


Part 1: Background and Motivation

Part 2: Methodology

Part 3: Experimental Results

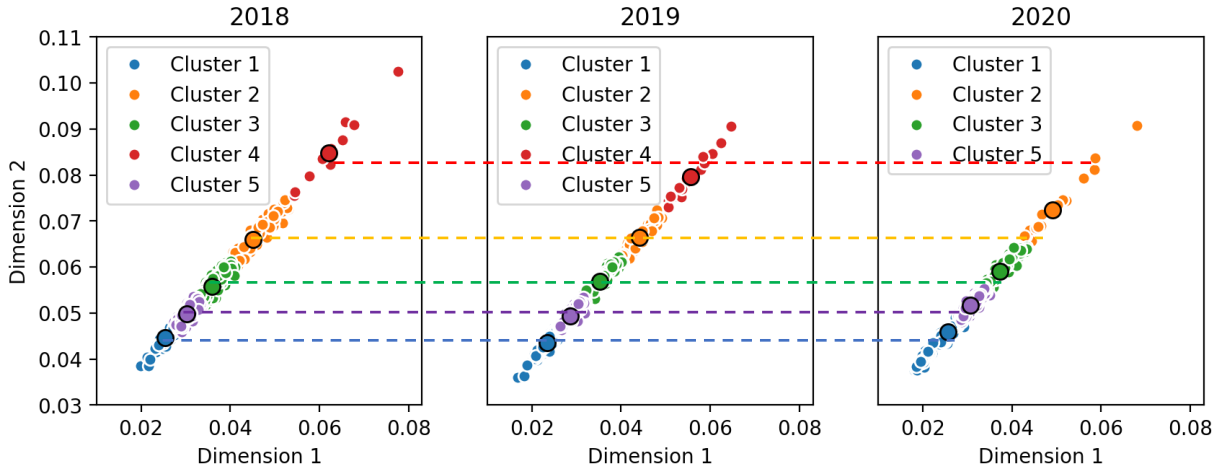
Visualization of VAE features



- The features of Friday (red) and Saturday (purple) are more concentration as the year increased, indicating less changes in the network traffic pattern happened on Friday and Saturday of 2020.
- The nighttime of the network traffic pattern is more concentration than the daytime.

* Nighttime ranges from 18:00 - 5:00 and daytime ranges from 6:00 - 17:00.

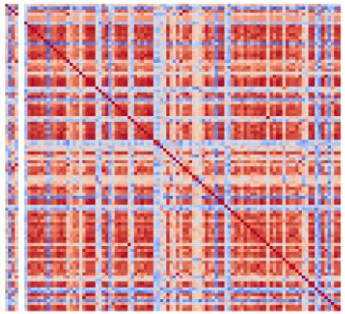
Clustering Results



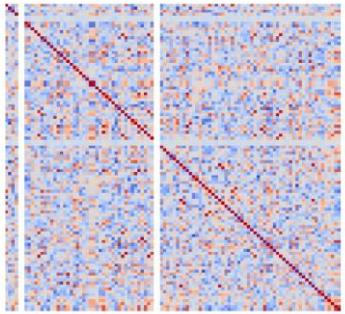
- The years 2018 and 2019 and year 2020 have different number clusters.
- The cluster 2 (orange) in 2018 and 2019 has a significantly lower position than it's in 2020.
- The year 2020 may happen some unique network traffic patterns as fewer employees are being physically in the office due to the work from home policies.

* The high-dimensional mean and standard deviation vectors are decomposed into 2-D dimension for visualization.
** the large circles referred to the cluster centers of different clusters

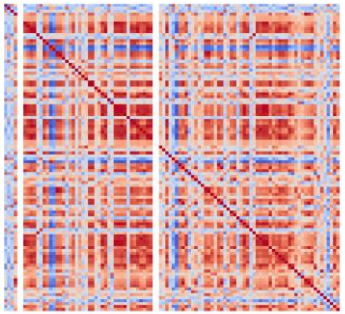
Anomalies



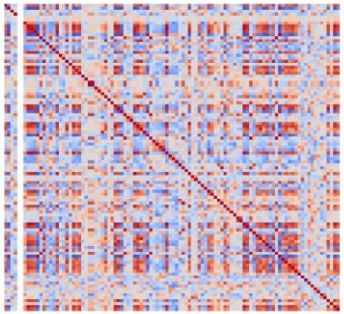
02/05/2020 nighttime



03/18/2020 nighttime



07/08/2020 daytime



05/24/2020 daytime

- While we can determine the anomalies by the purely data-driven approach, the implicit meaning of these anomalies is still unclear.
- A solid explanation may be found in understanding the reasons behind the anomalies based on physical knowledge of the network domain.

* Nighttime ranges from 18:00 - 5:00 and daytime ranges from 6:00 - 17:00.

Thank you!