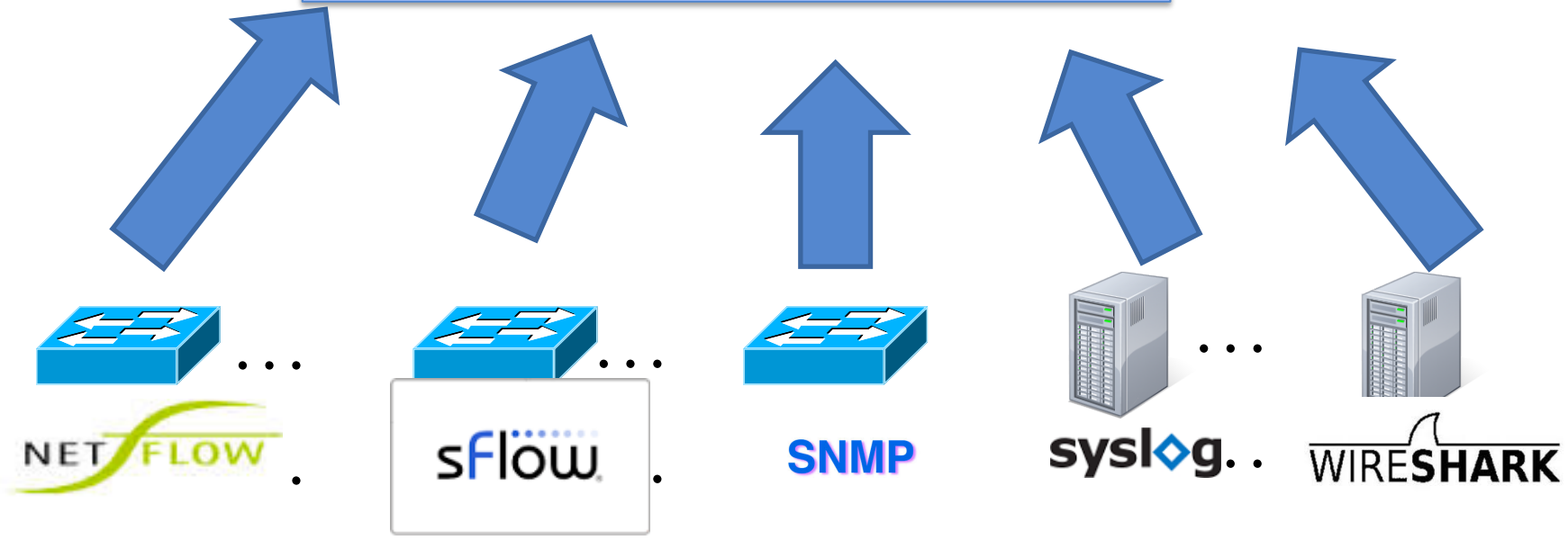
Data Analytics for Network Telemetry

Minlan Yu
Harvard University

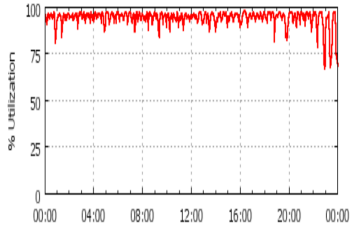
What is Network Telemetry?

Data analytics on the collector



Importance of Data Analytics for Network Telemetry

Performance



Utilization

Network Telemetry

Availability

Security

Challenges of Data Analytics for Network Telemetry

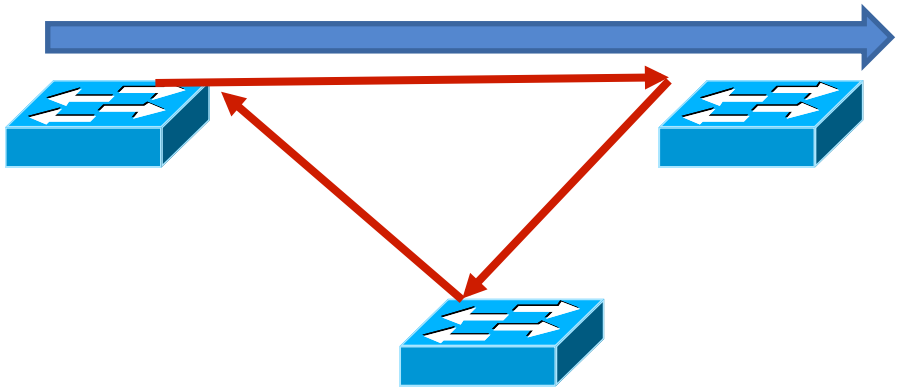
A variety of network-specific,
real-time analytics

Many detailed data
from heterogeneous sources

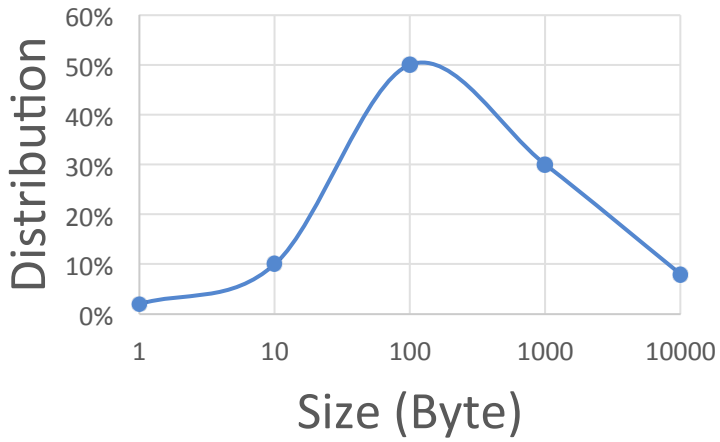
Challenge 1: Scalability to Many detailed Data

Every Flow Matters

Transient loop/blackhole



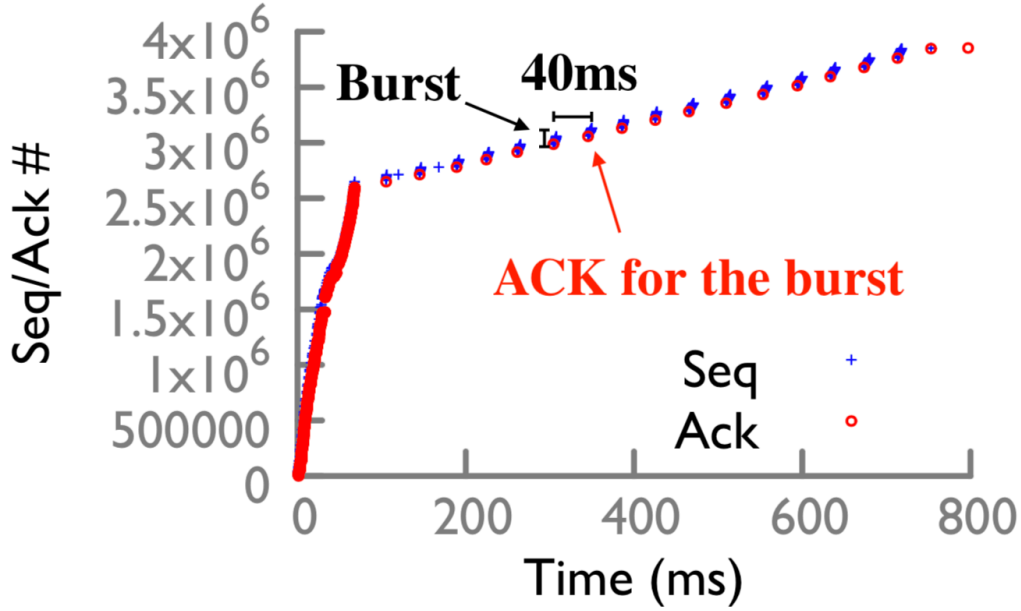
Fine-grained traffic analysis



- 10-100K flows per device * 10K-100K devices

Every Packet Matters

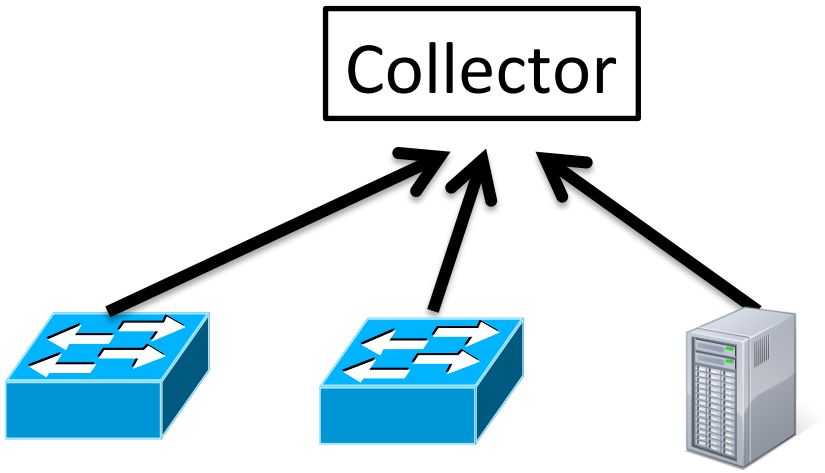
- Tail latency problems everywhere
  - Terasort 200 GB on 20 servers on EC2
  - 6.2K connections
  - Flows with 99.9 percentile latency
    - Delayed ACK, RTO, packet losses,
    - slow start, fast recovery etc.
  - Cannot predict which flow/packet sees which problem



- 10M packets per 10G port * 10K-1M ports

Solution: Improving Scalability

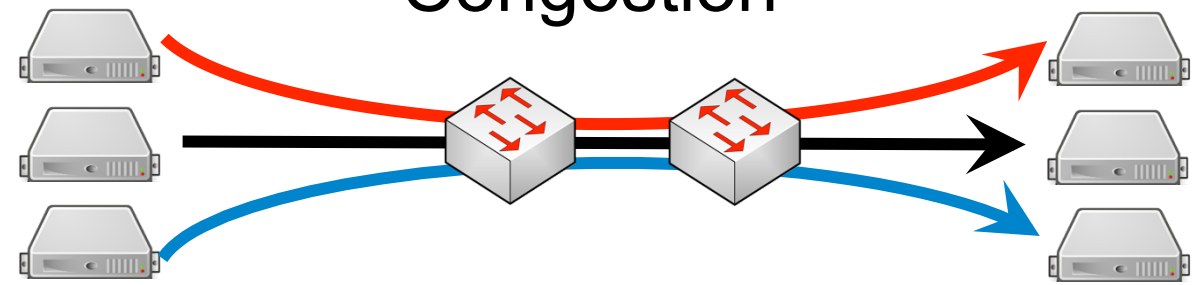
- Compress data with sketches
  - E.g., UniMon, FlowRadar, NitroSketch, SketchVisor, etc.
- Adapt data collection based on queries
  - E.g., OpenSketch, DREAM, EverFlow, Sonata etc.
- Challenge: Divide analytics between data sources and the collector
  - Limited processing speed relative to traffic rate
  - Limited network to transfer the data
  - Various programmability, computing, memory

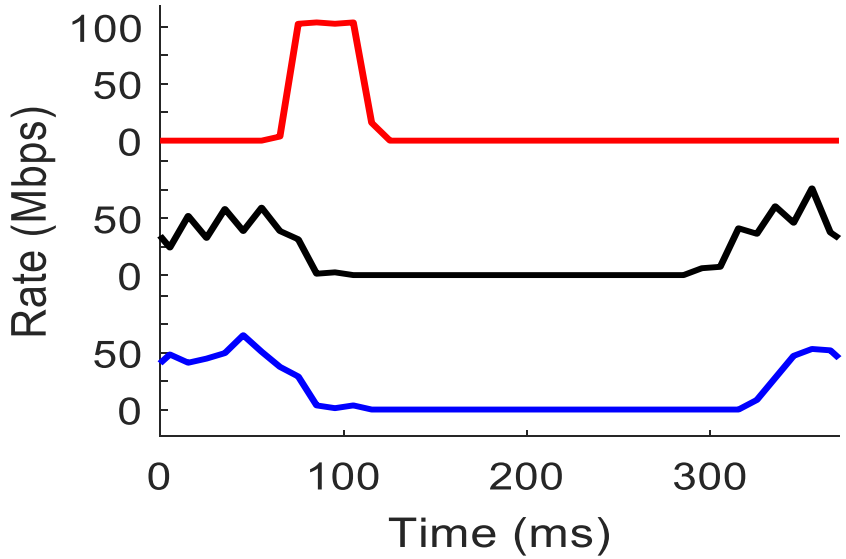


Challenge 2: Real-time

Capture Fine Time Scale Events

Detecting Transient Congestion



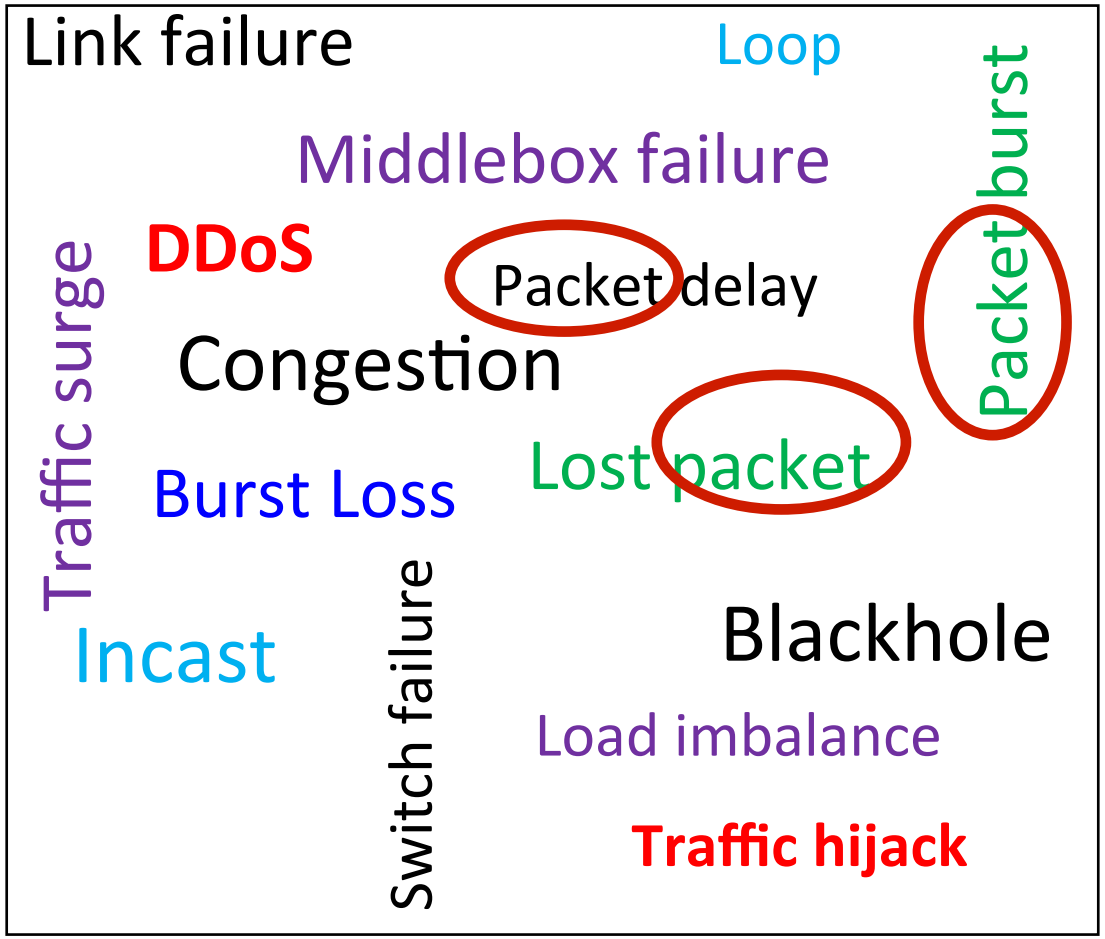


40 ms burst

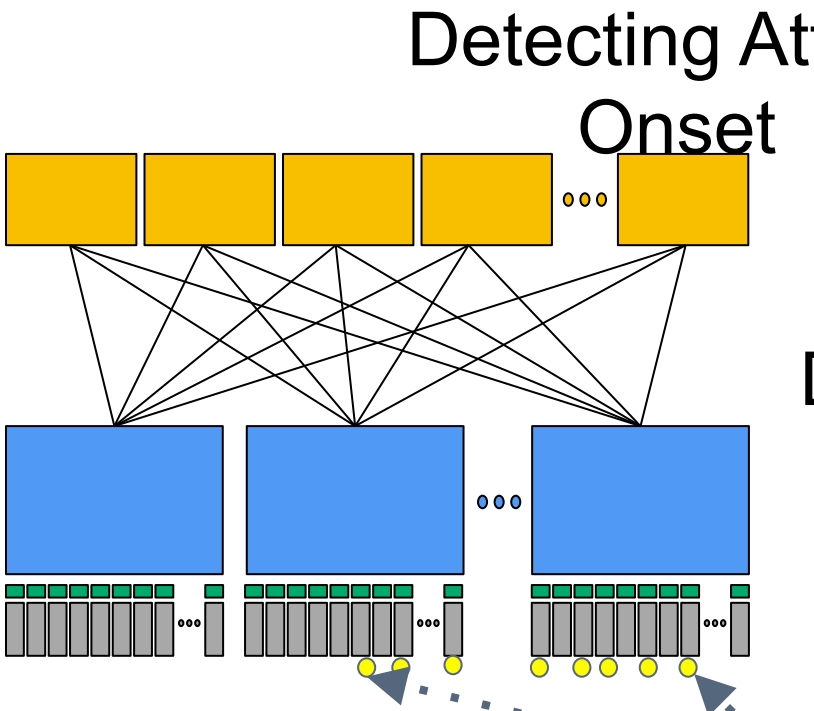
Timeouts lasting several 100 ms

Trumpet (SIGCOMM'16)

Packet-level Events in Sub-milliseconds



Fine Timescale Events across the Network



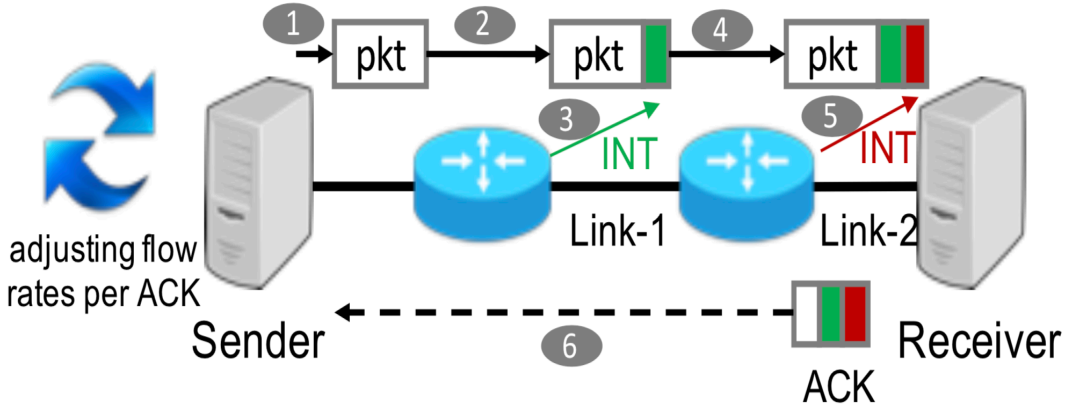
Trumpet (SIGCOMM'16)

Capture and Analyze Events in Real-time

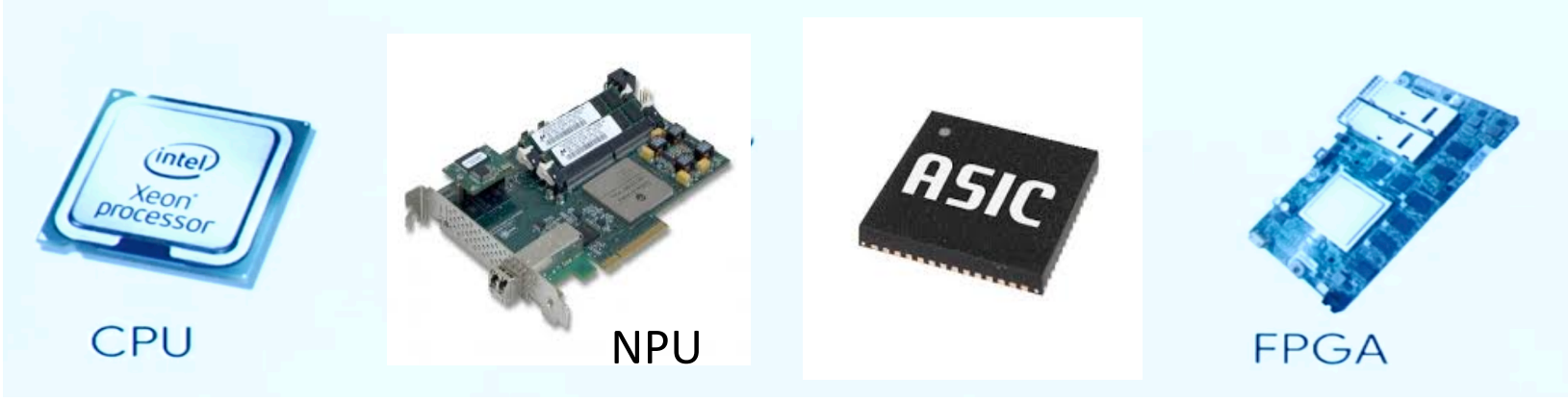
- Real-time detection means capital savings
  - A DDoS attack could cost an enterprise more than $2 million [Kaspersky Lab's IT Security Risks Survey]
  - AWS provides 30% refund for anything below 99.0% uptime
- Fast reaction to real-time events
  - Fast failure localization and recovery
  - Fast traffic engineering and congestion control

Solution: Analytics in Real-time

- In data plane control loop
  - In data plane event capturing: e.g., INT
  - In data plane prediction and reaction
- Challenge:
  - How to speed up analytics to sub-milliseconds
  - Or compile analytics down to the data plane



HPCC (SIGCOMM'19)



Challenge 3: Diverse Data Sources

Diverse Network Data in the Complex Networks

- Physical network
  - Servers: Pingmesh, NetBouncer, sFlow, etc.
  - Switches: SNMP, Syslog, NetFlow, packet traces, loss rate, interface counter
- Other network layers
  - Routing, traffic enginering, load balancing, firewalls
- Connecting to ISPs
  - Internet path availability, BGP, DNS
- Applications
  - Connectivity and performance logs in storage, database, ML etc.

Example: Incident Routing

- The curse of dimensionality
  - Need many training examples in proportion to #features
  - But incidents are rare events
- Diverse data formats
  - Data available at different components, regions
  - … with different frequency, scale, accuracy
- Limited visibility into each teams, especially in evolving networks
  - No one person can understand, parse, clean all the data
  - Yet, network components, monitoring data evolve all the time

Ongoing work with Microsoft

Solution: Handling Diverse Data Sources

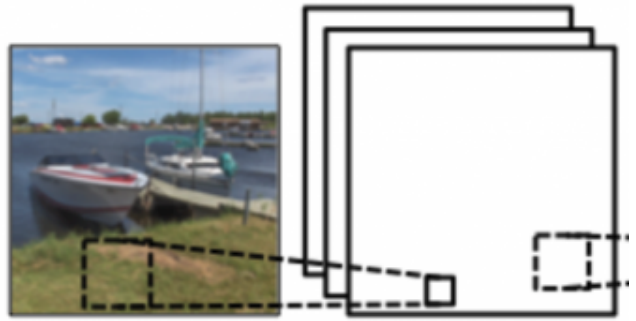
- Distributed, per-team predictors instead of global classifiers
  - Each team analyze if it should be involved in handling the incident
  - Fewer dimensions
  - Encode local data, local dependencies, local changes

- Challenge: Distributed analytics
  - How to divide problems, data sources, analytics?
  - For a broader set of problems

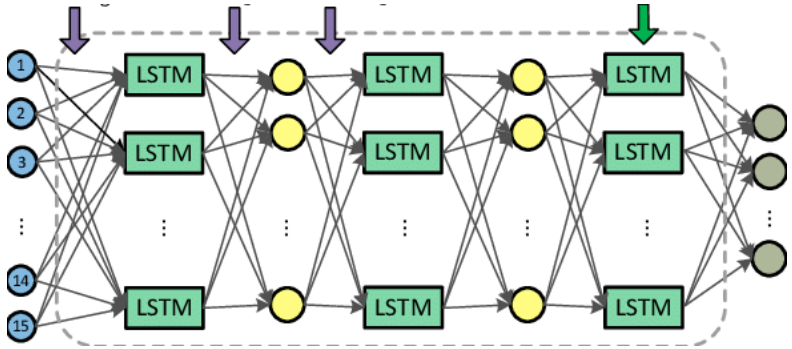
Challenge 4: Network Specific Analytics

Topological Information

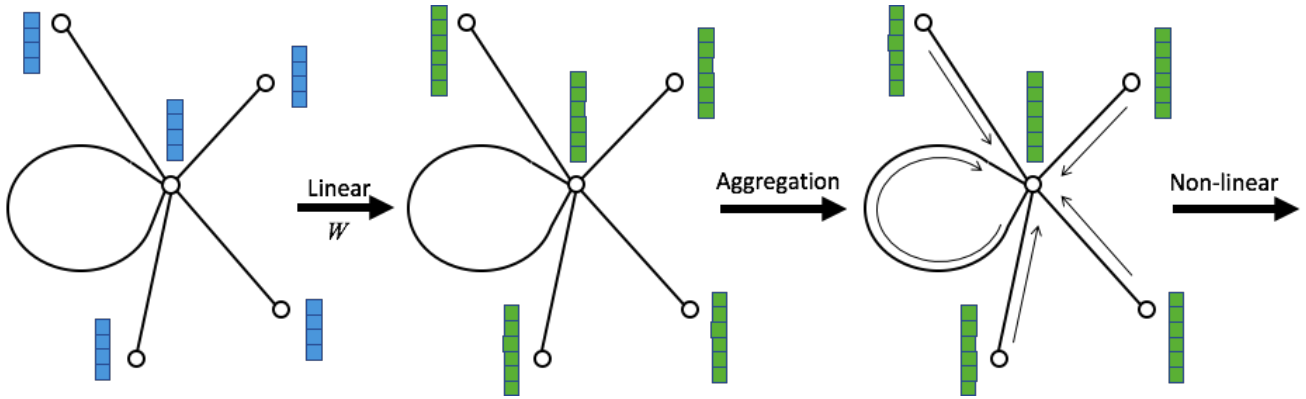
Spatial



Temporal



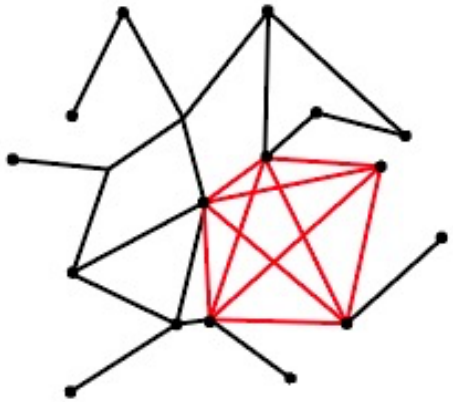
Topological



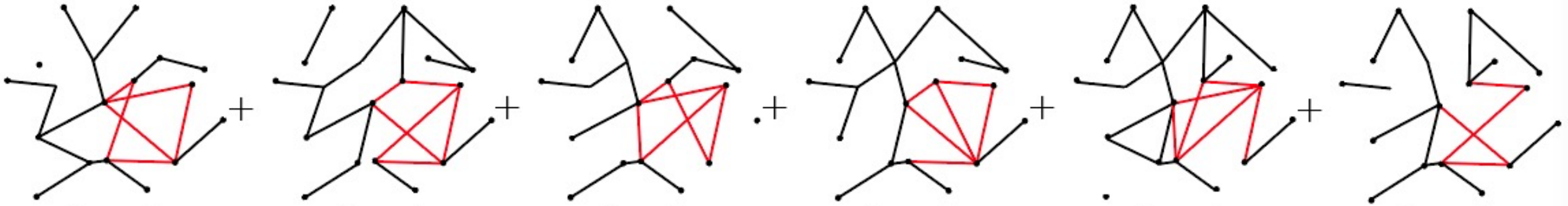
Dynamic Information

- P2P Botnet detection as an example

Static:

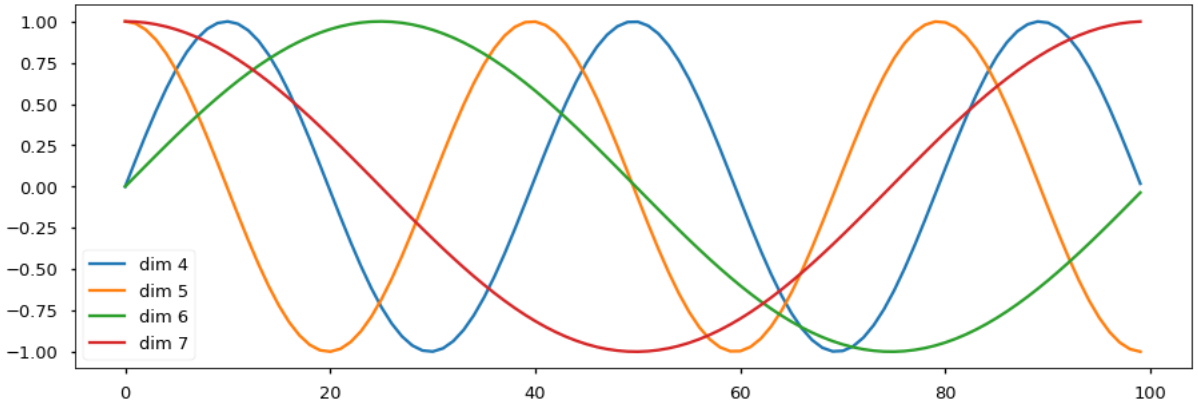


Dynamic:



Solution: Customized Analytics for Network Telemetry

- Our solution
  - Use graph convolutional neural network to encode topological information
  - Embed discrete timesteps using sinusoids of different periods
  - Aggregate features across time at each edge using LSTM



- Challenge: New analytics abstractions and frameworks for network-specific feature

Ongoing work

Summary of Challenges

- Scalability
- Real-time
- Diverse data sources
- Network specific data analytics

Network Telemetry and Analytics in Wide Area

- Scalability: Even larger scale of data from IoT devices
- Real-time: More variant, challenging networks
- Diverse data sources: More heterogeneity and sometimes mobile
- Network specific analytics: A broader set of queries

Thank you