Expectations to Machine Learning for Network Management

NML-RG meeting, IRTF/IETF 97 (Soeul)
November, 2016

Kohei Shiomoto (NTT)

Outline

- Network management issues (3min)
- Data-driven approach (3min)
- Two examples of our practice (9min)
  1. SYSLOG analytics
  2. Trouble Ticket analytics

     Explain only our goal and key idea

     Skip through several slides on math
- Discussion

Challenges in ICT Management

- Diversified applications & services/Modern Web traffic
  - Streaming, browsing, SNS, e-commerce, e-Health, …
- High expectations for Availability & Quality
  - 99.9999% availability, high resolution, low noise, stalling free, quick response, …
- Complex ICT system structure
  - Devices, software, protocols, …
- Interaction of players
  - Customer, ISPs, CDN, …
- Communications get encrypted. https://…
- …

Disaggregation -why we pursue?

- Vertically integrated system
  - Network devices (router, switch, middle-box, etc.) have been vertically integrated system.
  - Those vertically integrated systems consist of many hardware and software components provided by different component vendors.
  - The end of life of some components could risk the life of entire system.
  - Adding new functionalities is under control of system vendor.
- Disaggregation could solve the issues.

SDN & NFV

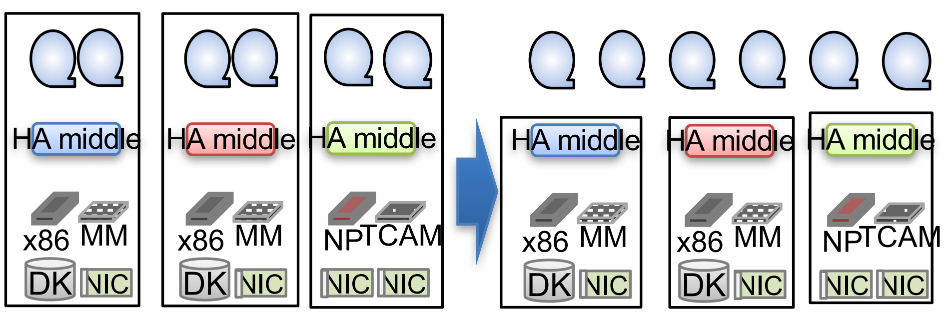
Software-driven Control

Software-Defined Networking (SDN)

- Routing & Signaling, traffic engineering/steering,
- Commodity L2/L3 Switch-Router hardware

Network Function Virtualization (NFV)

- Middlebox (NAT, FW, IDS/IPS, VPN, CDN, …)
- Commodity X86 machine hardware



Cost of disaggregation -Increase of complexity

- Disaggregation brings about a number of benefits: cost reduction, elastic capacity, quick deployment of new functions.
- But those benefits could not be obtained without sacrifice.
- Network device, which is disaggregated into components, introduces further complexity caused by interactions among components.
- Each component is frequently and quickly replaced with newer one as new features are developed and released.

How to deal with complex system?

- Powerful network management paradigm is required to deal with complexity.
- We could not rely on traditional mechanism-driven approach, which pieces together accurate mechanisms of individual components.
- We have to rely on a holistic data-driven approach to model an entire system by analyzing relationship between inputs and outputs.
  - *Conventional* Mechanism-driven approach
    - Given understanding precise mechanisms of components, build up a model of entire system.
  - *Towards* Data-driven approach
    - Given data, infer the relationship between inputs and outputs.
    - Machine learning is a key.

Varieties of data can be used.

- Traffic load
- Performance
- Syslog
- Trouble tickets
- SNS messages (e.g. Twitter)
- …

Numerical, text, …

SYSLOG Analytics

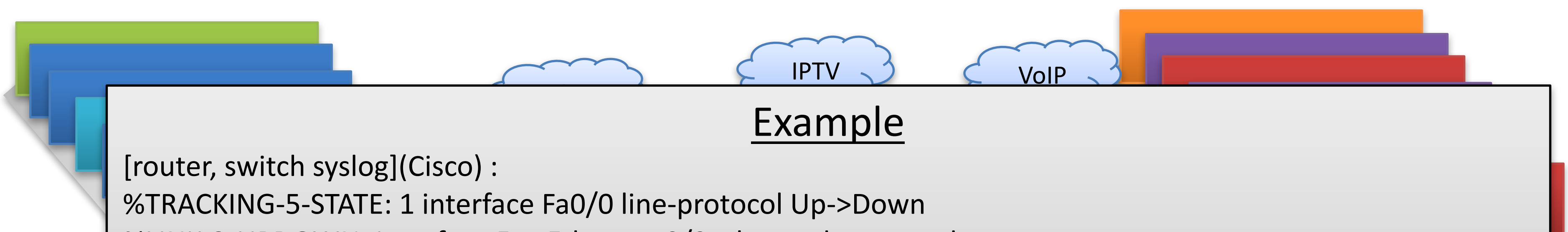
Innovative R&D by NTT

Spatio-Temporal Factorization of Log Data for Understanding Network Events

Tatsuaki Kimura, Keisuke Ishibashi, Tatsuya Mori,
Hiroshi Sawada, Tsuyoshi Toyono, Ken Nishimatsu,
Akio Watanabe, Akihiro Shimoda, Kohei Shiomoto

Background

- Various network logs are gathered by NMSs and monitored
  - Switch, router, RADIUS sever,…
  - syslog, server log, alarm, SNMP trap, …
  - logs contain useful information for NW trouble shooting



Example

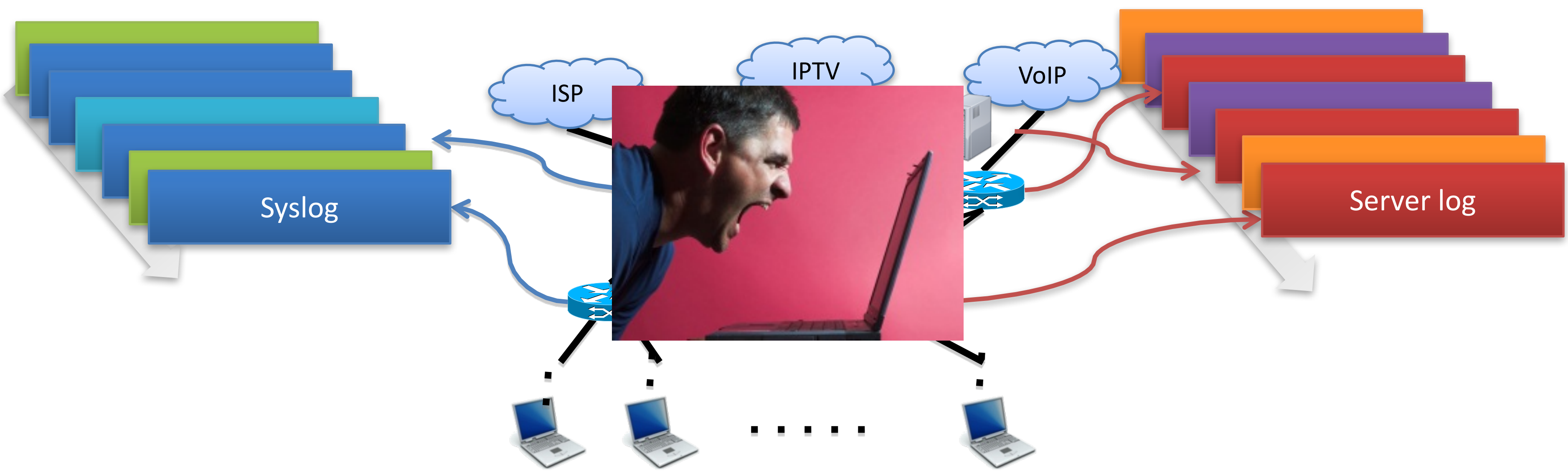
[router, switch syslog](Cisco) :
%TRACKING-5-STATE: 1 interface Fa0/0 line-protocol Up->Down
%LINK-3-UPDOWN: Interface FastEthernet 0/9, changed state to down
%SYS-5-CONFIG I: Configured from console by vty2 (10.11.11.11)

[NMS alarm]:
[エラータイプ 100]リンクダウンが起こりました. ホスト名:hostA IPアドレス: 10.11.1.1 プロトコル3
[エラータイプ 100] パケットロスを検出しました. type: xxxx, host: hostB, IPアドレス: 10.11.11.11
<優先度: 低> オペレータがログインしました. from yyyy, host: hostC, 2011/11/11

Background

- Various network logs are gathered by NMSs and monitored
  - Switch, router, RADIUS sever,…
  - syslog, server log, alarm, SNMP trap, …
  - logs contain useful information for NW trouble shooting
- Diverse and massive amounts of logs
  - multiple venders, multiple services, complex network events
  - over 1,000,000 of messages/day

$\Rightarrow$ *Analyzing logs has become serious problem*



Our research goal

Mining *network (NW) event information* from large and diverse NW log data

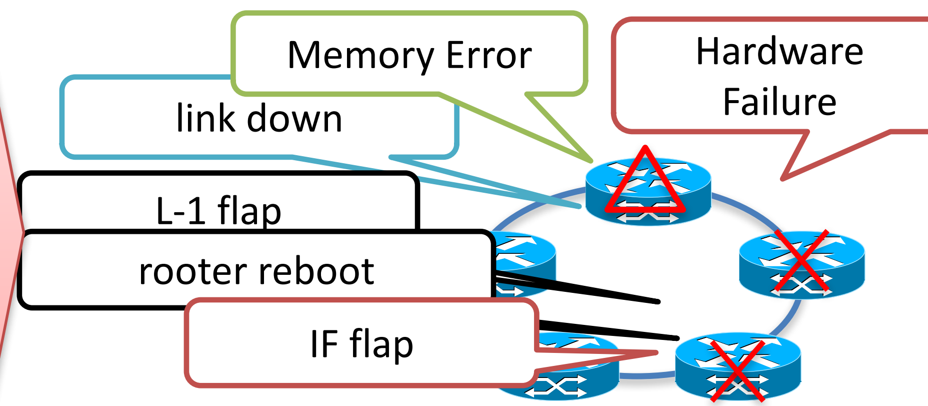
Network events = *spatial and temporal patterns of log messages*

messages associated with initialization of various process caused by router reboot event

multiple layer flaps caused by L-1 flap (L-2, OSPF re-convergence, BGP flap, ...)

virtual path dis-connection related to physical machine failure

```
2012-1-1T00:00:00 %TRACKING-5-STATE: 1 interface Fa0/0 line-protocol Up->Down
2012-1-1T00:00:00 %LINK-3-UPDOWN: Interface FastEthernet 0/9, changed state to
down
2012-1-1T00:00:00 %SYS-5-CONFIG I: Configured from console by vty2 (10.11.11.11)
2012-1-1T01:11:00 msg [100]:  STP: VLAN 1 Port 38 STP State -> DISABLED (PortDown)
2012-1-1T01:11:00 msg [101]: System: Interface ethernet 38, state down
2012-1-1T03:00:00 msg [200] : STP: VLAN 100 Port 22 STP State -> DISABLED
(PortDown)
2012-1-1T03:00:00 msg [201] : System: Interface ethernet 22, state down
2012-1-1T00:00:00 %SYS-5-CONFIG I: Configured from console by vty2 (10.11.11.1
2012-1-1T10:30:00 System: Interface ethernet 1, state down
2012-1-1T10:30:00 System: Interface ethernet 1, state up
2012-1-1T10:30:00 System: Interface ethernet 2, state down
2012-1-1T12:00:00 init: alarm-control (PID 111) terminate signal sent
2012-1-1T12:00:00 init: bslockd (PID 124 ) terminate signal sent
2012-1-1T12:00:00 init: ce-l2tp-service (PID 123 ) terminate signal sent
2012-1-1T12:00:00 init: chassis-control (PID 1111 ) terminate signal sent
2012-1-1T12:00:00 init: disk-monitoring (PID 7082 ) terminate signal sent
2012-1-1T00:00:00 %SYS-5-CONFIG I: Configured from console by vty2 (10.11.11.1
2012-1-1T15:45:10 msg [200] : STP: VLAN 100 Port 22 STP State -> DISABLED
(PortDown)
2012-1-1T15:45:10 msg [201] : System: Interface ethernet 22, state down
2012-1-1T16:12:40 System: Interface ethernet 1, state down
2012-1-1T16:12:40 System: Interface ethernet 1, state up
2012-1-1T16:12:40 System: Interface ethernet 2, state down
2012-1-1T20:30:00 init: alarm-control (PID 111) terminate signal sent
2012-1-1T20:30:00 init: bslockd (PID 124 ) terminate signal sent
2012-1-1T20:30:00 init: ce-l2tp-service (PID 123 ) terminate signal sent
2012-1-1T20:30:00 init: chassis-control (PID 1111 ) terminate signal sent
```

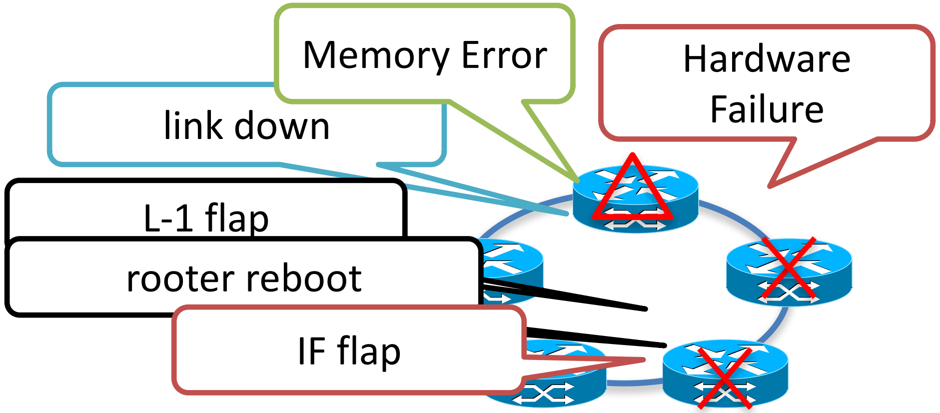


Why *mining NW events*?

Automatically constructing domain knowledge of NW operators
(without requiring skills and experience)

Many possible applications:

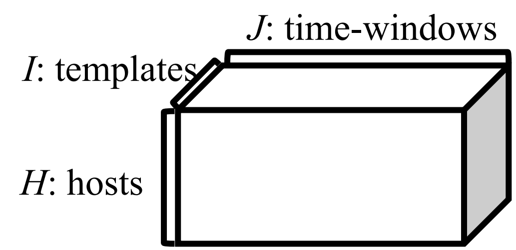
- Obtaining new alarm rules
- Quick understanding of root cause of problems
- May help in detection of "silent failure"



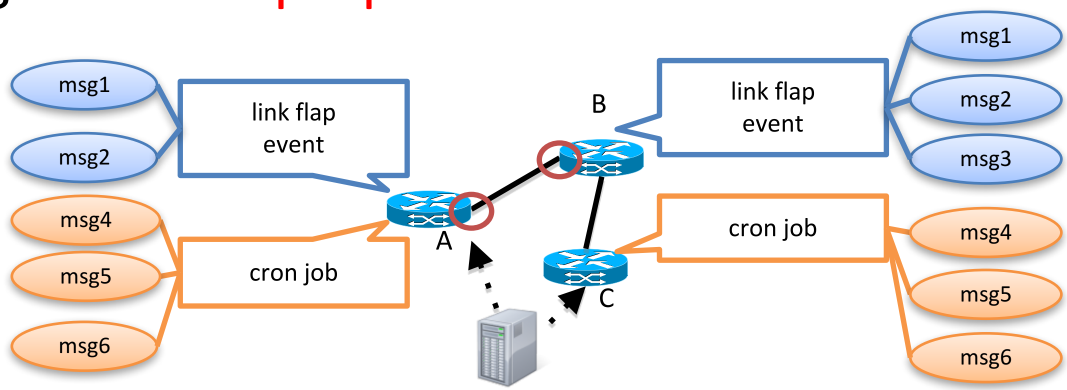
Key idea (observation)

Log data can be considered as rank-3 tensor

types of log messages (templates), hostname, time



Observed log data = 'superposition' of NW events



	msg1	msg2	msg3	msg4	msg5	msg6
A	○	○		△	△	△
B	○	○	○			
C				△	△	△

Extraction of NW events problem ⇒ Tensor Factorization problem

Challenges & Summary of Contributions

[1] Unstructured and massive log messages

More than 1,000,000 lines/day

Formats of log messages depend on vendor or service

⇒ We present **Statistical Template Extraction (STE)**

- automatically extract primary templates from large log data

[2] Log data are very complex

Underlying network events occur all around network

Network events span across several locations, network layers, and services

⇒ We present **Log Tensor Factorization (LTF)**

- extract *spatial and temporal patterns* of log data
- based on Nonngegative Tensor Factorization (NTF) approach

STE Motivation

Raw log messages cannot be used directly

- log messages contain various *parameters* (IP address, host name, PID,…,etc.)
- To correlate log messages, we need to know log *templates* = messages without parameters

```
%TRACKING-5-STATE: 1 interface Fa0/0 line-protocol Up->Down
%LINK-3-UPDOWN: Interface FastEthernet 0/9, changed state to down
%SYS-5-CONFIG I: Configured from console by vty2 (10.11.11.11)
```

```
%TRACKING-5-STATE: * interface * line-protocol Up->Down
%LINK-3-UPDOWN: Interface FastEthernet *, changed state to down
%SYS-5-CONFIG I: Configured from console by * (*)
```

STE Methodology

1. **Scoring frequency of words** among similar messages

    parameter words appear infrequently compared to template words in each position

2. **Clustering score**, and determine *parameter words* for each message

    thresholds for score of *parameter* words differ depending on log messages

    density-based clustering algorithm (***DBSCAN)***

raw log messages:

<189> security telnet connection 15720 with 10.7.11.11 broken
<189> security telnet connection 18340 with 10.8.9.123 broken

1	2	3	4	5	6	7
<189>	security	telnet	15720	with	10.7.11.11	broken
<189>	security	telnet	18340	with	10.8.9.123	broken

Remove

log template:

<189> security telnet connection * with * broken

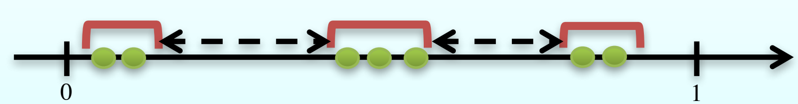
①Scoring:

If *word* appears in *P*-th position in log that contains *L* words:

$$Score(word, P, L) = \Pr(word \mid P, L)$$

②Clustering scores (DBSCAN):

Distance between each cluster is > δ



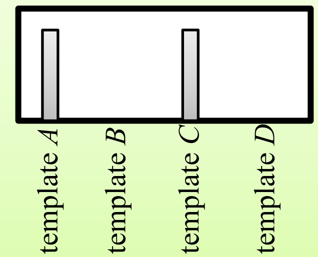
Spatial & temporal patterns we want to extract

Hierarchical correlations are observed in log data

Definition 1 [Template Group]

- group of templates that tend to co-occur: $l = (i_1, i_2, \ldots)$
  - ✓ represents ***event at individual host***
  - ✓ e.g. linkflap: linkdown + linkup

    reboot: process initialization messages

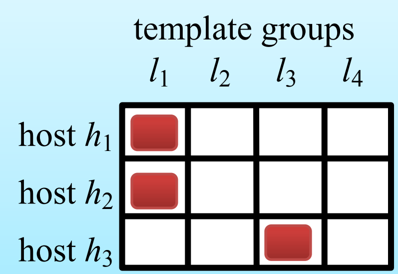


Definition 2 [Network Event]

- set of tuples <host, template groups> that tend to co-occur

$$e = \{(h_1, l_1), (h_2, l_2), \ldots\} \quad (h_1, h_2 \in H, \quad l_1, l_2 \in L)$$

  - ✓ ***spatial extension*** of template groups
  - ✓ e.g. link down event among neighboring hosts



LTF

Log Tensor $\boldsymbol{X}$ $(I \times H \times J)$

x_{ihj}: # of occurrences of template i at host h, time window j

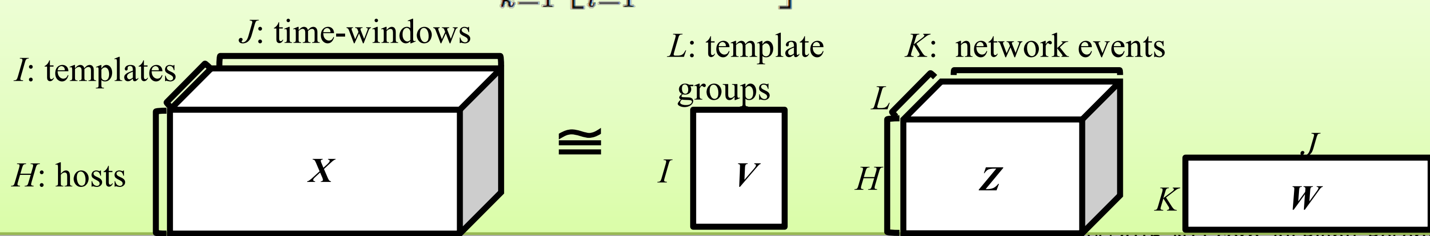
- templates: $1,\ldots, i,\ldots, I$
- hosts: $1,\ldots, h,\ldots, H$
- time windows: $1,\ldots, j,\ldots, J$ (log data are partitioned)

- LTF factorizes $\boldsymbol{X}$ into

1. Log template matrix $\boldsymbol{V} = [\boldsymbol{v}_l]$ $(I \times L)$
2. Network event tensor $\boldsymbol{Z} = [z_{lk}]$ $(L \times K \times H)$
3. Weight matrix $\boldsymbol{W} = [\boldsymbol{w}_k]$ $(K \times J)$

* K: # of network events. L: # of template groups (given)

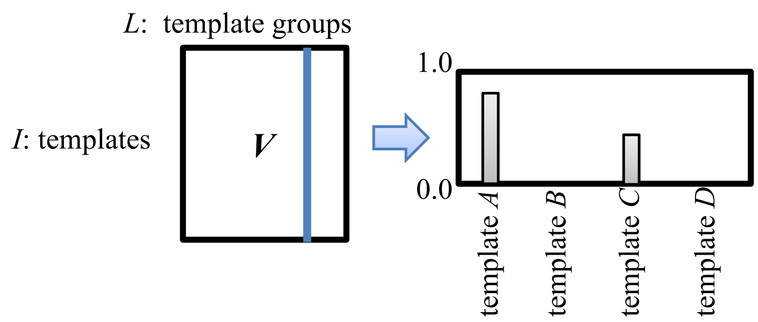
$$\boldsymbol{X} \simeq \sum_{k=1}^{K} \left[\sum_{l=1}^{L} \boldsymbol{v}_l \otimes \boldsymbol{z}_{lk} \right] \otimes \boldsymbol{w}_k,$$



Intuitive interpretations of LTF

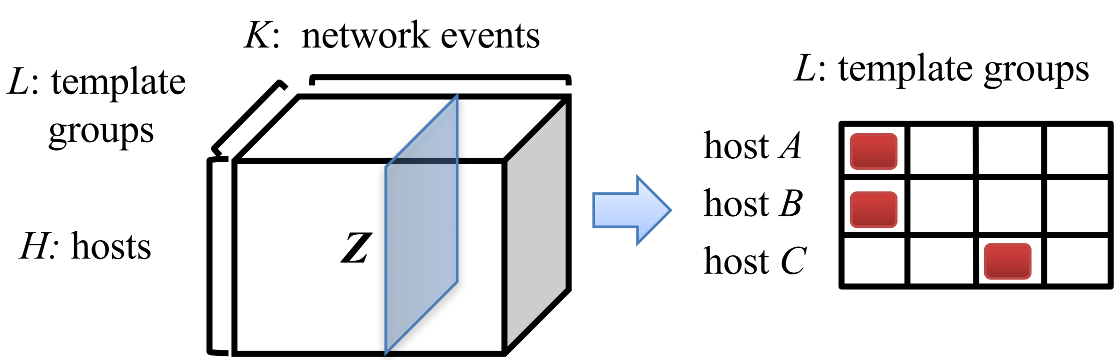
template group matrix V

l-th row of matrix V represents l-th template group
⇒ corresponds to Template Group



network event tensor Z

k-th slice of Z represents which template groups occur at which hosts
⇒corresponds to Network Event



Formulation & Algorithm

LTF Problem Formulation

Optimization problem with nonnegative constraints on each tensor

Objective function = KL-divergence between $\boldsymbol{X}$ and $\boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{W}$

LTF Problem

$$\min_{\boldsymbol{V},\boldsymbol{Z},\boldsymbol{W}} \mathcal{D}(\boldsymbol{X} \| \boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{W}),$$

$$\text{s.t.} \quad \boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{W} \geq \boldsymbol{O}, \ \sum_i v_{il} = 1, \sum_{l,h} z_{lkh} = 1,$$

$$\mathcal{D}(\boldsymbol{X} \| \boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{W}) = \sum_{i,h,j} x_{ihj} \log \frac{x_{ihj}}{\sum_{k,l} v_{il} z_{lkh} w_{kj}} - x_{ihj} + \sum_{k,l} v_{il} z_{lkh} w_{kj}$$

Algorithm (multiplicative update rules; type of ***EM algorithm***)

simple and iterative form

$$v_{il} := \frac{\sum_{h,j,k} \frac{\breve{z}_{lkh} \breve{w}_{kj}}{\sum_{k',l'} \breve{v}_{il'} \breve{z}_{l'k'h} \breve{w}_{k'j}} \cdot x_{ihj}}{\sum_{h,j,k} z_{lkh} w_{kj}} \breve{v}_{il},$$

$$z_{lkh} := \frac{\sum_{i,j} \frac{\breve{v}_{il} \breve{w}_{kj}}{\sum_{k',l'} \breve{v}_{il'} \breve{z}_{l'k'h} \breve{w}_{k'j}} \cdot x_{ihj}}{\sum_{i,j} v_{il} w_{kj}} \breve{z}_{lkh},$$

$$h_{kj} := \frac{\sum_{i,h,l} \frac{\breve{v}_{il} \breve{z}_{lkh}}{\sum_{k',l'} \breve{v}_{il'} \breve{z}_{l'k'h} \breve{w}_{k'j}} \cdot x_{ihj}}{\sum_{i,h,l} v_{il} z_{lkh}} \breve{w}_{kj},$$

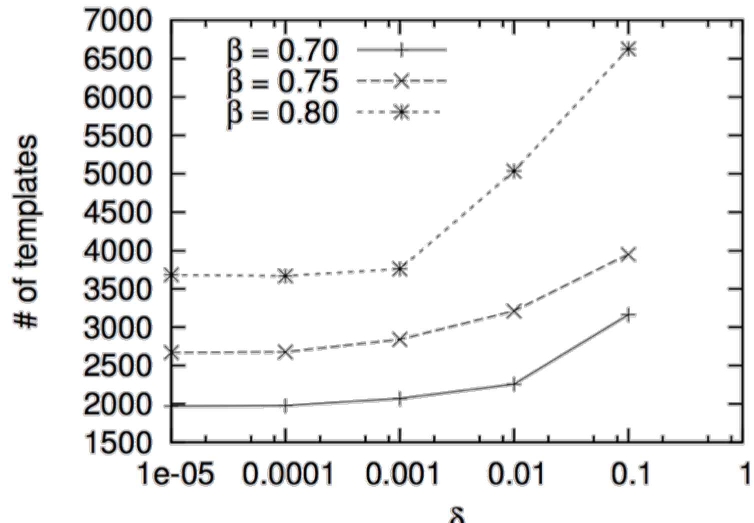
STE Evaluation

Data Set

- 5 million lines of logs (captured in small network, 5 months)

Evaluation metrics = effectiveness

- calculate # of **extracted templates**
- 5,000,000 logs ⇒ 2,000 ~ 3, 000 templates (less than 1%)



False positive words ⇒ 0.07 ~ 0.08% of 1,000,000 words

LTF Evaluation Metrics

Data Set

over 600,000 lines of 1-day log data

dimensions are roughly $100\ (I) \times 150\ (H) \times 150\ (J)$

Evaluation Metrics

expressive power: How well does LTF fit to real data?
$\Rightarrow$ use well known measure '**average test log-likelihood**'

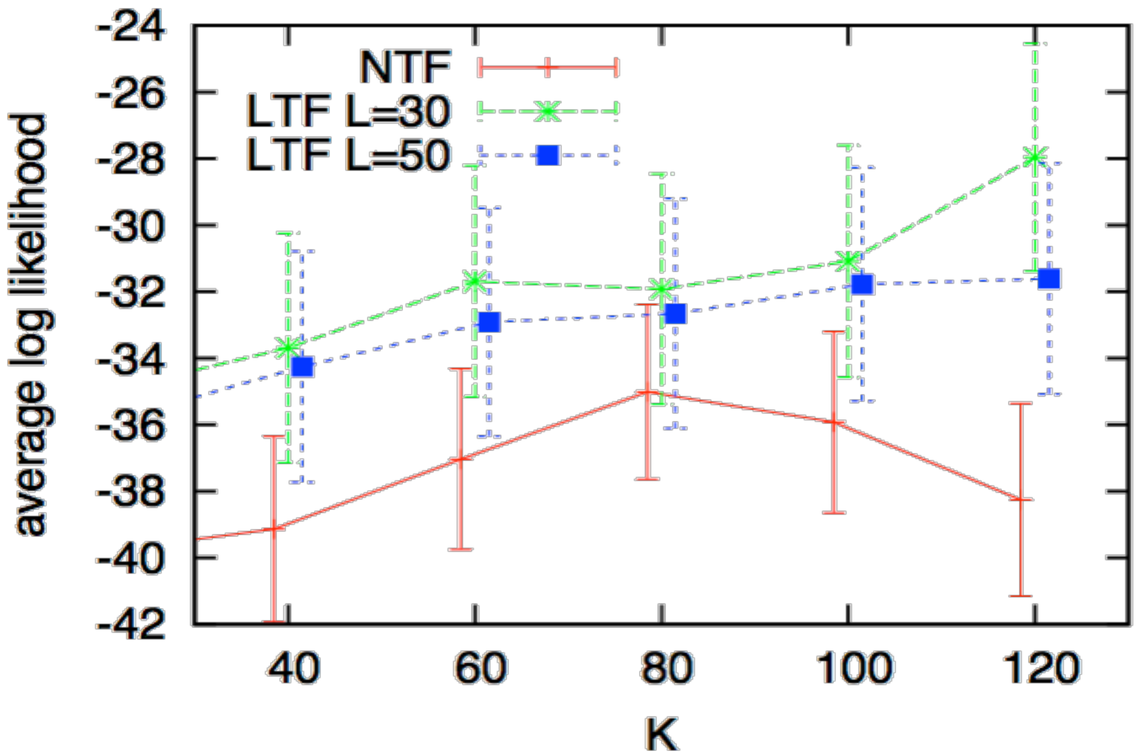
- prediction power of randomly masked elements
- higher value means better fit to data

LTF Evaluation Results

average test log-likelihood with different K and L

we used normal NTF model as baseline

* L: # of template groups, K: # of network events



LTF fits better to real data than current NTF

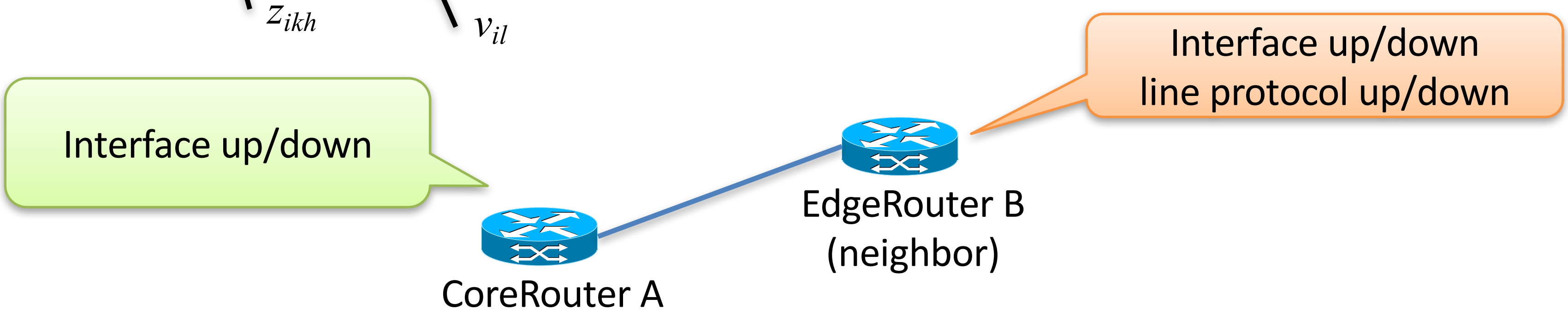
Case study results (Examples of output of LTF) 1/2

Neighboring link flap event

NEIGHBORING LINK FLAP EVENT

Host Name	TG Weights	Weights	Templates
CoreRouterA	0.666	0.4	TIME : ifmgr [*] : %PKT_INFRA-LINK-3-UPDOWN : Interface * , changed state to Up
		0.4	TIME : ifmgr [*] : %PKT_INFRA-LINK-3-UPDOWN : Interface * , changed state to Down
		0.2	TIME : %SYS-3-LOGGER_DROPPED : System dropped * console debug messages.
EdgeRouterB	0.333	0.4	* : * : %LINK-3-UPDOWN : Interface * , changed state to up
		0.17	* : TIME : %LINK-3-UPDOWN : Interface * , changed state to administratively down
		0.17	* : * : %LINEPROTO-5-UPDOWN : Line protocol on Interface * , changed state to up
		0.17	* : * : %LINEPROTO-5-UPDOWN : Line protocol on Interface * , changed state to down
		0.05	* : * : %LINK-3-UPDOWN : Interface * , changed state to down

z_{ikh} v_{il}

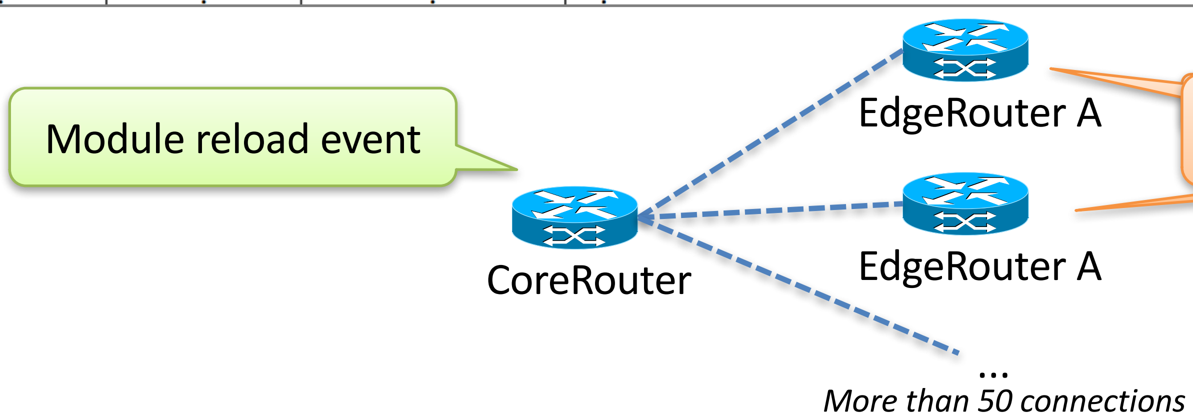


Case study results (Examples of output of LTF) 2/2

Tunneling path disconnection event

TUNNELING PATH DISCONNECTION EVENT

Host Name	TG Weights	Template Weights	Templates
CoreRouter	0.0253	0.4375	SNMP Trap: a status change for a module. Software image for the module is missing or invalid....
		0.0833	os: loader: * for * is *
		0.0833	id of requester is *
		0.0833	OsCrashDump: invalid crash record skipped, ...
		⋮	⋮
EdgeRouterA	0.01656	0.92805	Tunneling Virtual Path * is disconnected, hardware unavailable.
EdgeRouterA	0.01594	0.98308	Tunneling Virtual Path * is disconnected, hardware unavailable.
⋮	⋮	⋮	⋮



Summary of SYLOG Analytics

Presented STE

- Extracting primary templates from noisy log messages
- Compressing 5,000,000 lines to 3000 templates

Presented LTF

- Modeling generation of logs as rank-3 tensor and factorizing into template groups and network events
- Much better than current NTF
- Can correctly extract hidden complex network events

Trouble Ticket Analytics

Innovative R&D by NTT

Workflow Extraction for Service Operation using Multiple Unstructured Trouble Tickets

Akio Watanabe, Keisuke Ishibashi, Tsuyoshi Toyono, Tatsuaki Kimura, Keishiro Watanabe, Yoichi Matsuo, Kohei Shiomoto

NTT Network Technology Laboratories

April 25, NOMS 2016

Background

- We would like to specify *troubleshooting process*
  - MTTR strongly depends on the time until deciding the resolution

<u>Troubleshooting procedure for system management</u>

- Why process is not defined?
  - various process for thousands of failures
  - requiring tacit knowledge
  - including domain rule

Present understanding of process

- Operators search the resolution from *trouble tickets*
  - amount of valuable knowledge about failures
- Much information are written by *natural language*

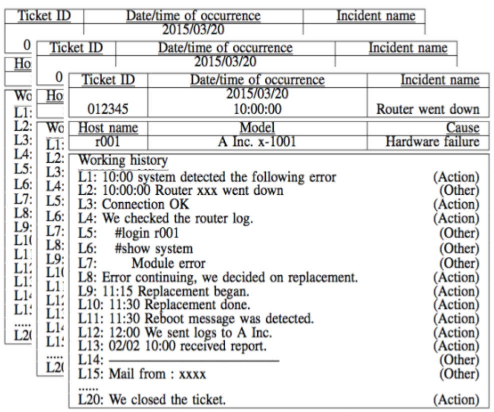
Ticket ID	Date/time of occurrence	Incident name
012345	2015/03/20 10:00:00	Router went down

Host name	Model	Cause
r001	A Inc. x-1001	Hardware failure

Working history	
L1: 10:00 system detected the following error	(Action)
L2: 10:00:00 Router xxx went down	(Other)
L3: Connection OK	(Action)
L4: We checked the router log.	(Action)
L5: #login r001	(Other)
L6: #show system	(Other)
L7: Module error	(Other)
L8: Error continuing, we decided on replacement.	(Action)
L9: 11:15 Replacement began.	(Action)
L10: 11:30 Replacement done.	(Action)
L11: 11:30 Reboot message was detected.	(Action)
L12: 12:00 We sent logs to A Inc.	(Action)
L13: 02/02 10:00 received report.	(Action)
L14: ————————————	(Other)
L15: Mail from : xxxx	(Other)
......	
L20: We closed the ticket.	(Action)

Our goal

- Automatically specifying <u>troubleshooting process</u> from trouble tickets
- Generating *Workflow*, ***graphical*** flowchart of actions for each failure

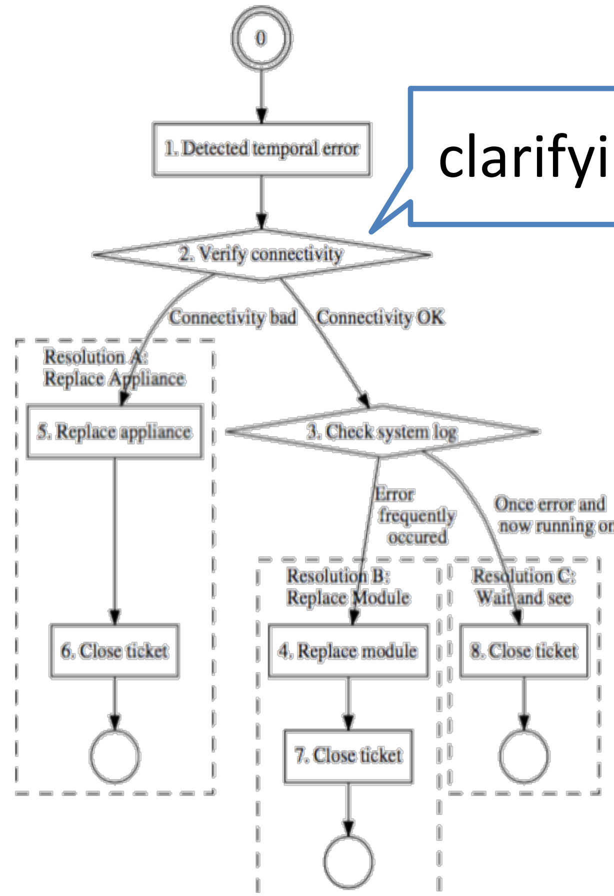


Ticket ID	Date/time of occurrence	Incident name
012345	2015/03/20 10:00:00	Router went down

Host name	Model	Cause
r001	A Inc. x-1001	Hardware failure

Working history

L1: 10:00 system detected the following error (Action)
L2: 10:00:00 Router xxx went down (Other)
L3: Connection OK (Action)
L4: We checked the router log. (Action)
L5: #login r001 (Other)
L6: #show system (Other)
L7: Module error (Other)
L8: Error continuing, we decided on replacement. (Action)
L9: 11:15 Replacement began. (Action)
L10: 11:30 Replacement done. (Action)
L11: 11:30 Reboot message was detected. (Action)
L12: 12:00 We sent logs to A Inc. (Action)
L13: 02/02 10:00 received report. (Action)
L14: ———————— (Other)
L15: Mail from : xxxx (Other)
......
L20: We closed the ticket. (Action)

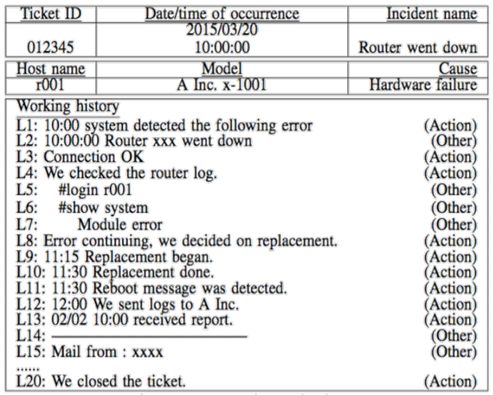


clarifying decide rules

summarizing multiple trouble tickets

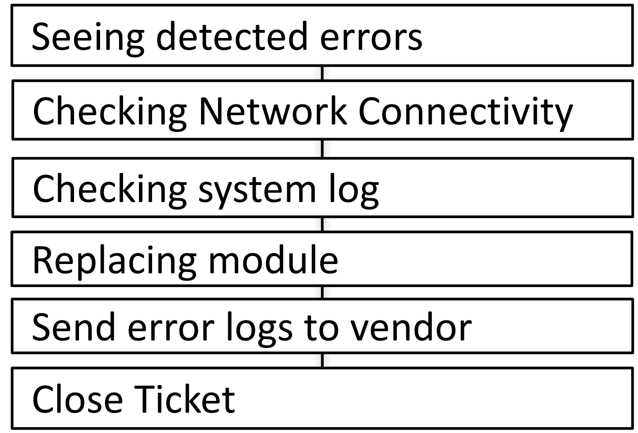
Two challenges

- Finding *action sequences* for each trouble ticket

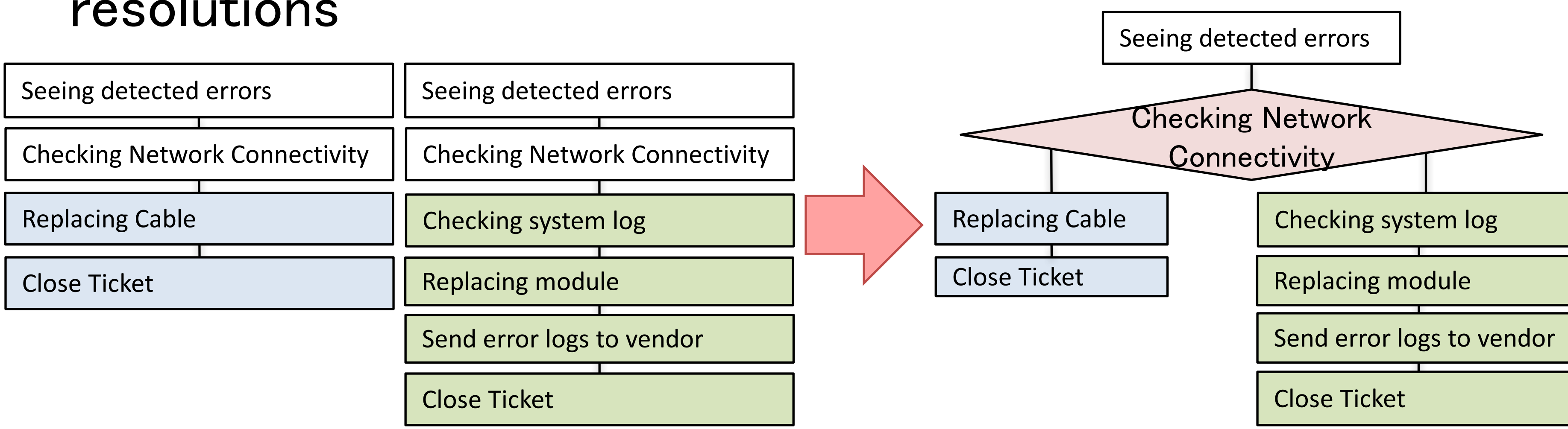


Ticket ID	Date/time of occurrence	Incident name
012345	2015/03/20 10:00:00	Router went down
Host name	Model	Cause
r001	A Inc. x-1001	Hardware failure

Working history
L1: 10:00 system detected the following error (Action)
L2: 10:00:00 Router xxx went down (Other)
L3: Connection OK (Action)
L4: We checked the router log. (Action)
L5: #login r001 (Other)
L6: #show system (Other)
L7: Module error (Other)
L8: Error continuing, we decided on replacement. (Action)
L9: 11:15 Replacement began. (Action)
L10: 11:30 Replacement done. (Action)
L11: 11:30 Reboot message was detected. (Action)
L12: 12:00 We sent logs to A Inc. (Action)
L13: 02/02 10:00 received report. (Action)
L14: ———————— (Other)
L15: Mail from : xxxx (Other)
......
L20: We closed the ticket. (Action)

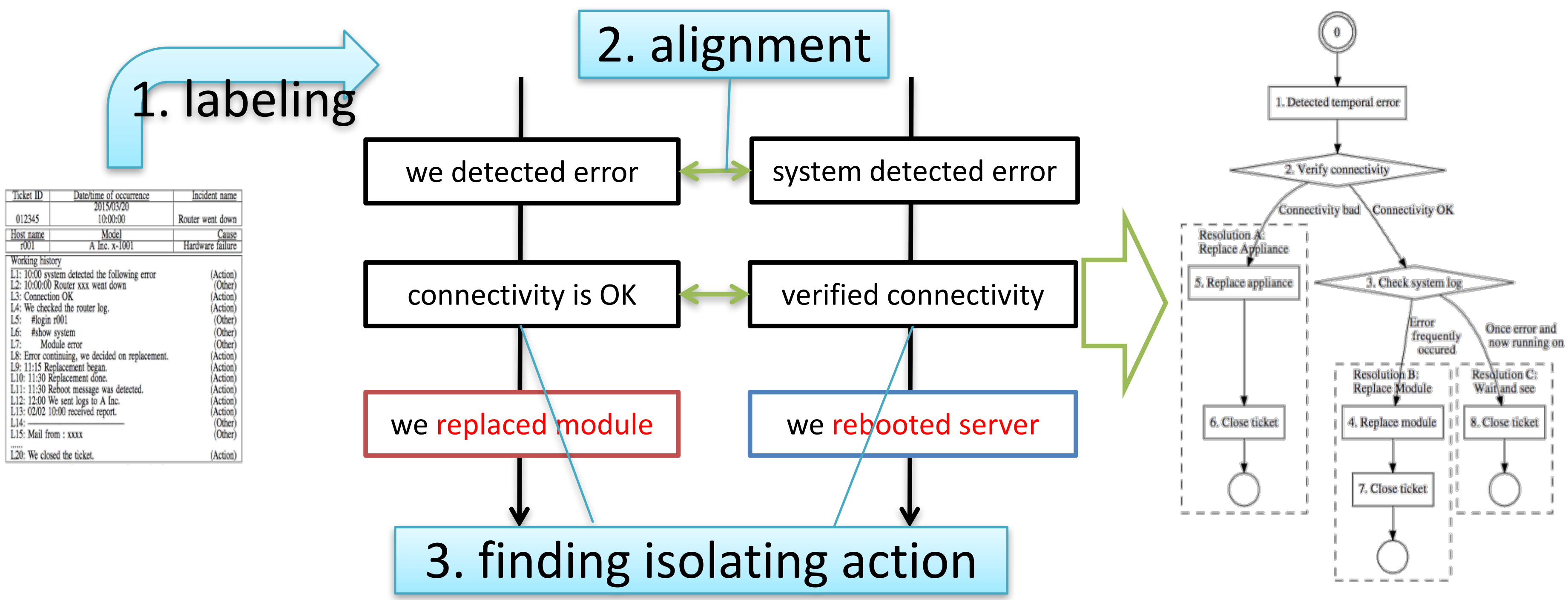


- Finding *isolating action;* that have multiple transitions to resolutions



Approach overview

- <u>3 steps</u> for extract workflow from multiple trouble tickets
  - 1. extract only action sentences from trouble tickets
  - 2. align the same messages in different tickets
  - 3. find operational change as a branch



1. Action Sentence Labeling

- Extracting sentences about (operator/system's) actions
  - Append sentences to labels indicating if is written about actions or not
- Supervised learning from labeled texts
  - Naive Bayes are used as classifier
  - Character-2gram is used as features

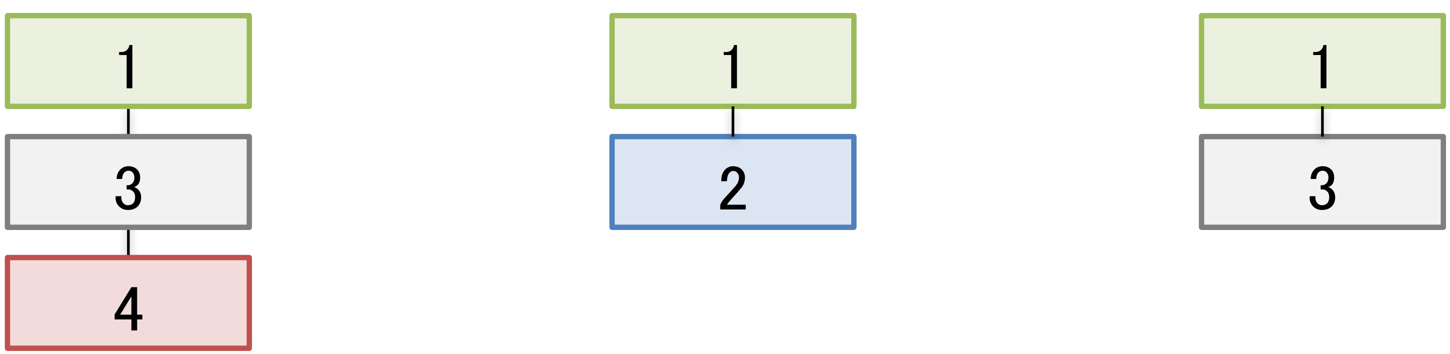
Action	10:00 Management system detected the following error
Other	10:00:00 Router xxx went down
Action	Connection Okay
Action	We checked the router log.
Other	#login r001
Other	#show system
Other	Error countinuing, we decided on replacement.
Action	11:15 Replacement began.
Action	11:30 Replacement done.
Action	11:30 Reboot message was detected.
Action	12:00 We sent system logs to A Inc.
Action	2/2 10:00 received report
Other	Mail from ...

2. Action Alignment

- Aligning sentences describing the same action

	Action sentences 1	Action sentences 2	Action sentences 3
1	System detected temporary error	Error is detected.	Error: Node down
2		We verified connectivity.	
3	Ping is OK		Ping NG
4	Many errors were in log	There is no error in log	

consider aligned numbers as action sequence



Formulation as *maximal matching Problem*

- Corresponding similar sentences
  - maximize the sum of similarities of aligned sentences

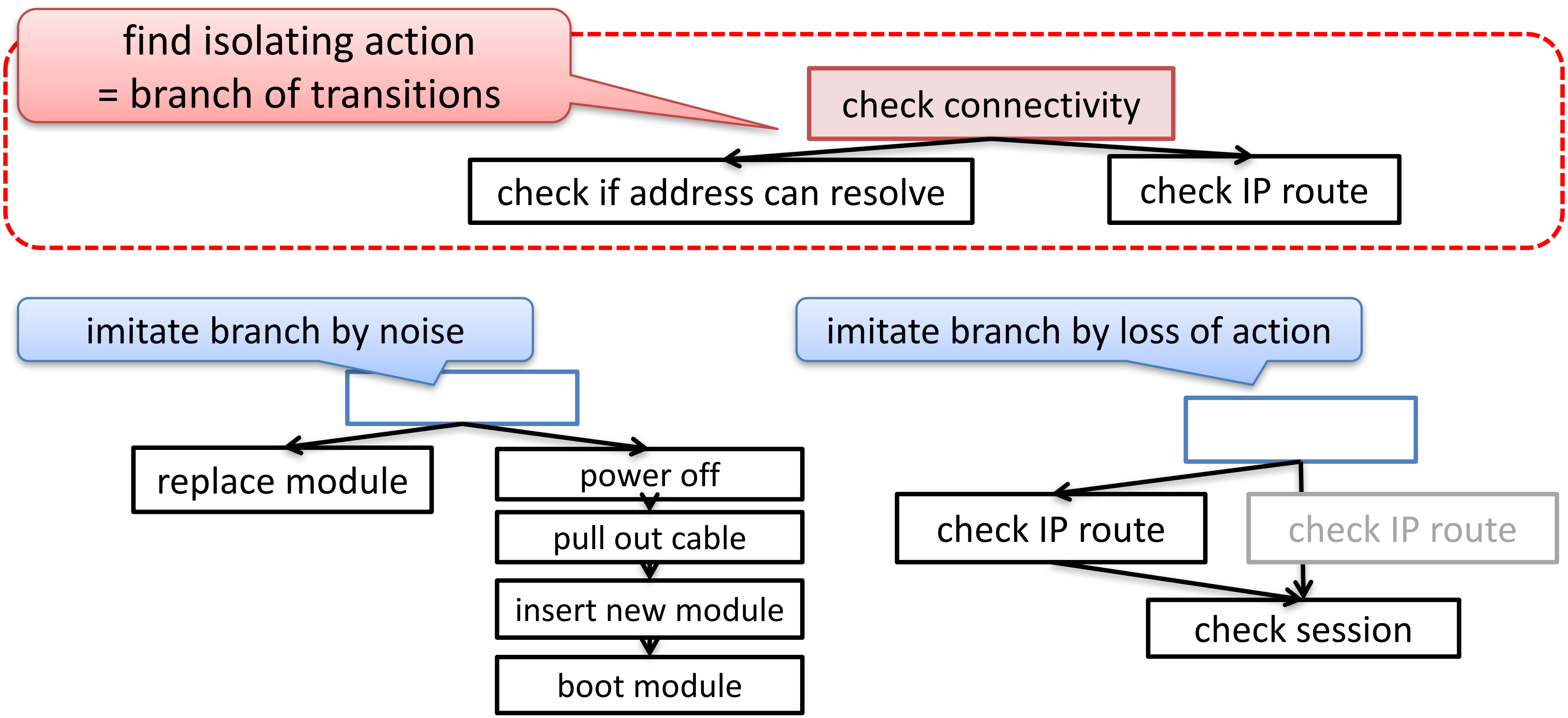
$$\hat{\mathbf{G}} = \arg\max_{\mathbf{G}} \sum_{j=1}^{J} \sum_{\substack{i,i' \in \mathcal{I} \\ i \neq i'}} \mathrm{sim}(G_{ij}, G_{i'j})$$

- Solving by *multiple sequence alignment* (MSA) method
  - MUSCLE algorithm [18] is used for aligning multiple sequences of sentences
  - dice coefficient as similarity of sentence pair

$$\mathrm{dice}(s_{ij}, s_{i'j'}) = \frac{2|s_{ij} \cap s_{i'j'}|}{|s_{ij}| + |s_{i'j'}|}$$

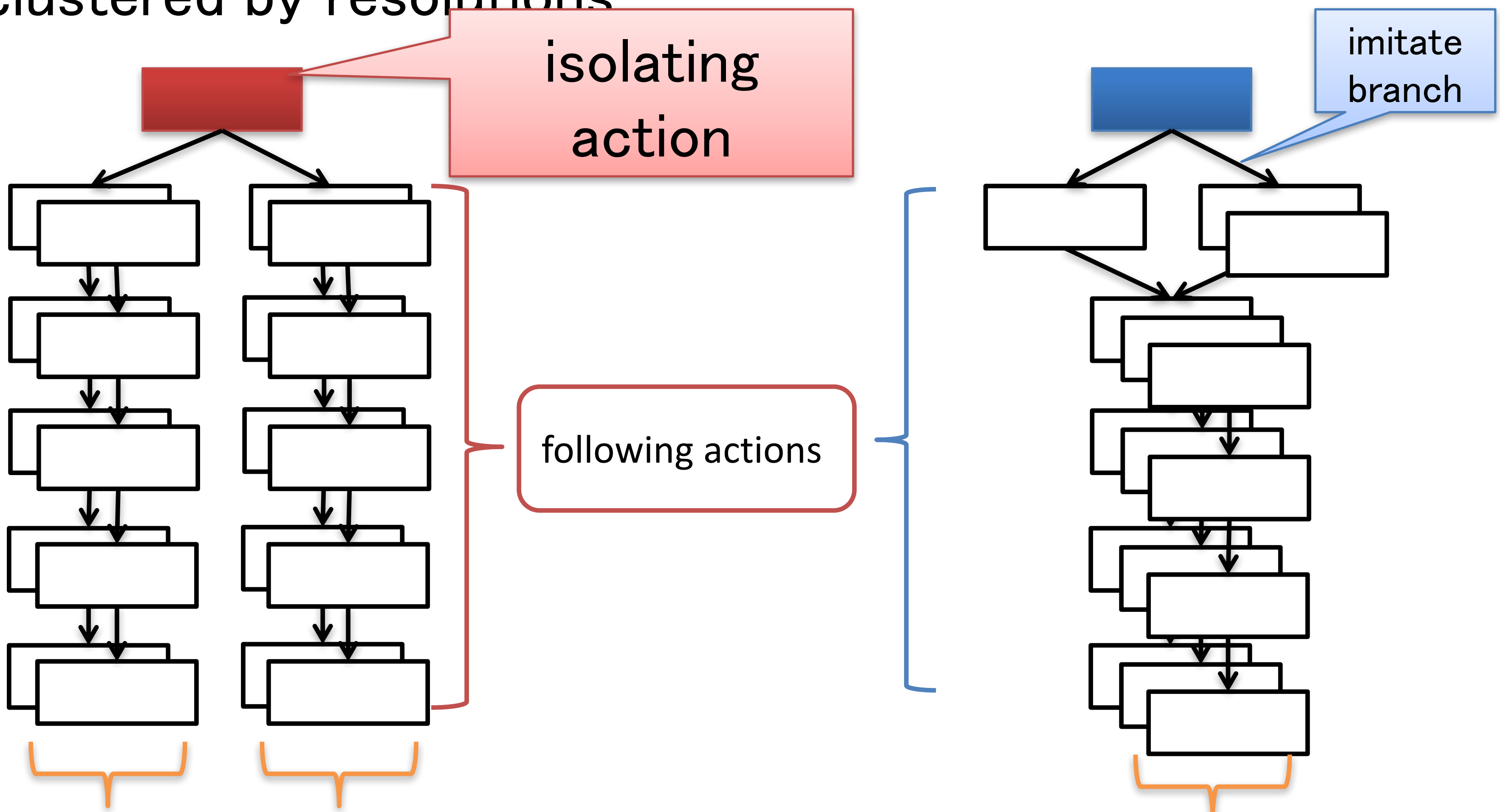
3. Isolating Action Searching

- Finding isolating action i.e. *branch of transitions to two resolutions*
- Problem: many imitate branches caused by loss or noise in action sequence



Key observation

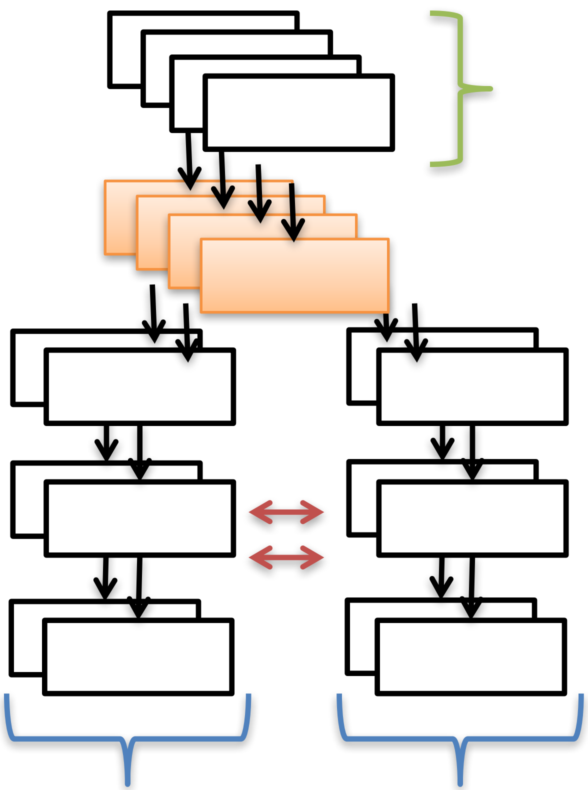
- the following actions of isolating action can be clustered by resolutions



seems to have two kinds of resolution

Method for finding isolating action

- 1. dividing the following actions using Spectral Clustering for each actions
- 2. choosing the action that has the best clustering score



similarity of clustered sentences

$$\mathrm{coh}(\mathbf{G}^{(\pm)}) = \sum_{i,i' \in \mathcal{I}^{(\pm)}} \left(|\mathbf{G}^{(\pm)}_{\{i\},\mathcal{J}}| - \mathrm{dist}(\mathbf{G}^{(\pm)}_{\{i\},\mathcal{J}}, \mathbf{G}^{(\pm)}_{\{i'\},\mathcal{J}})\right).$$

dissimilarity of not clustered sentences

$$\mathrm{dev}(\mathbf{G}^{(+)}, \mathbf{G}^{(-)}) = \sum_{i^+ \in \mathcal{I}^{(+)}} \sum_{i^- \in \mathcal{I}^{(+)}} \mathrm{dist}(\mathbf{G}_{\{i^+\},\mathcal{J}}, \mathbf{G}_{\{i^-\},\mathcal{J}}).$$

Score of the action

$$\begin{aligned}\mathrm{Score}(\mathbf{G}, \mathcal{I}^{(+)}, \mathcal{I}^{(-)}, j) &= \mathrm{coh}(\mathbf{G}_{\mathcal{I},\mathcal{J}\setminus\mathcal{J}_{j\prec}}) \\ &+ \mathrm{coh}(\mathbf{G}_{\mathcal{I}^{(+)},\mathcal{J}_{j\prec}}) + \mathrm{coh}(\mathbf{G}_{\mathcal{I}^{(-)},\mathcal{J}_{j\prec}}) \\ &+ \mathrm{dev}(\mathbf{G}_{\mathcal{I}^{(+)},\mathcal{J}_{j\prec}}, \mathbf{G}_{\mathcal{I}^{(-)},\mathcal{J}_{j\prec}}).\end{aligned}$$

Experiments

- Dataset: practical trouble tickets for network system
  - Written by Japanese
    - primitive linguistic preprocessing are executed
  - Separated into subsets by detected error

	the number of tickets	resolutions / of tickets
(i)	5	(A) turn on breaker (x2) (B) wait & see (x3)
(ii)	4	(A) replace module (x2) (B) replace port interface (x1) (C) replace cable (x1)
(iii)	3	(A) power outage (x2) (B) replace ONU (x1)
(iv)	29	(A) wait & see (x12) (B) detail log analysis (x15) (C) replace module (x2)

- given parameter
  - the threshold of similarity for alignment
  - the number of isolating actions

Quantitative evaluation

- We compared obtained result with ground truth
- Comparison of
  - alignment result & manually appended action ID
  - clustering result & manually checked true resolution for each ticket
  - words of extracted isolating actions & isolating action in true operation document

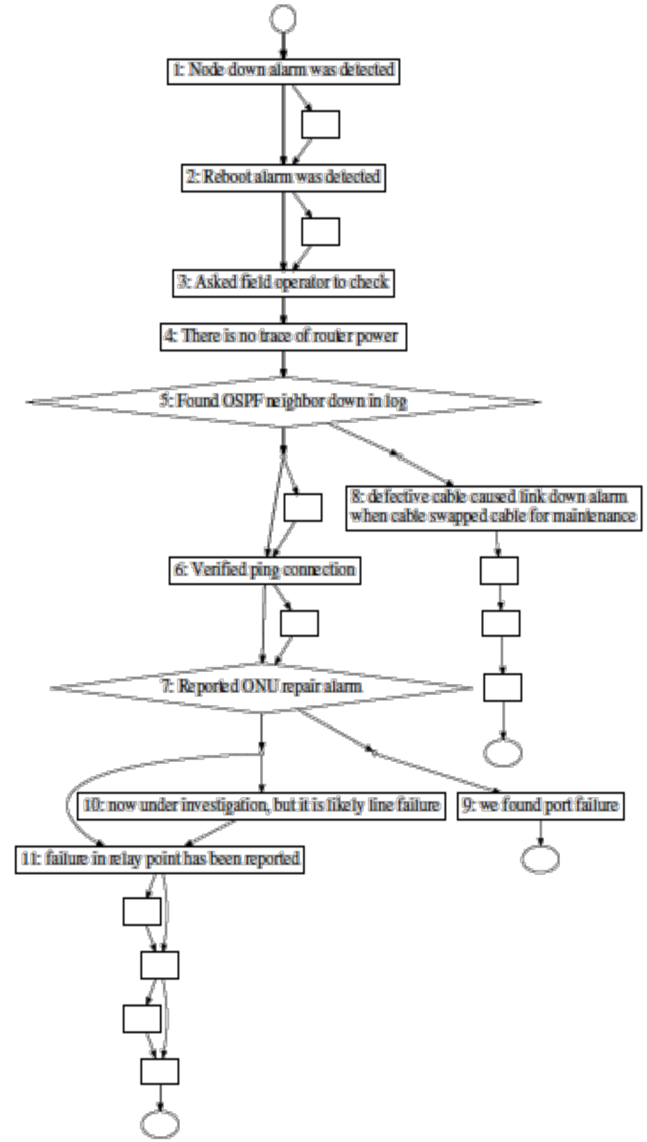
set	tickets(I)	Precision/Recall	isolating actions	extracted isolating actions	resolutions	clusters
(i)	5	87.1%/79.6%	check ONU power off	ONU/power/off/trace	(A)turn on breaker, (B)wait and see	{A,A},{B,B,B}
(ii)	4	87.7%/84.4%	(a)check OSPF down log, (b)field check	(a)OSPF/Neighbor/down, (b)repaired/Alarm	(A)replace module, (B)replace port, (C)replace cable	{A,A},{B},{C}
(iii)	3	94.9%/78.9%	field check	field/check	(A)power outage, (B)replace ONU	{A,A},{B}
(iv)	29	83.5%/70.2%	(a)check if temporal error by tool (b)reboot was reoccured	confirm/Module/Fault/transition temporal/error	(A)wait and see, (B)detail log analysis, (C)replace module	$\{A \times 12, B \times 3\}$, $\{B \times 12\}$, $\{C, C\}$

- miss of alignment is limited the case in
  - exchange of order
- miss of isolating action searching is caused by
  - the loss of isolating action
  - multiple (three or more) isolations

Case study result

- frequent actions are the same with the actions in manual document
- causes are described into the next actions of isolating actions

ID	description	resolution
1	node down alarm was detected	
2	reboot alarm was detected	
3	asked field operator to check	
4	there is no trace of router power off.	
5	found OSPF neighbor down message in log	
6	verified ping connection.	
7	reported ONU repair alarm	
8	defective cable caused link down alarm when cable swapped for maintenance.	wait until maintenance is over
9	we found port failure.	wait and see
10	now under investigation, but it is likely line failure.	replacement
11	failure in relay point has been reported.	replacement



Summary of Trouble Ticket Analytics

- Proposed extracting method for workflow automatically from multiple trouble tickets
  - Action Sentence Labeling
  - Action Alignment
  - Isolating Action Searching
- Future works
  - relaxing the limitation of order of actions
  - finding multiple isolating actions

Expectations to Machine Learning

- Correlation and Causality Inference
- Anomaly Detection
- Root Cause Analysis
- Knowledge Discovery

Diagnosing & Trouble-shooting

- Prediction
- Detection
- Root Cause Analysis
- Recovery

Concluding remarks
- Data-driven approach

- Disaggregate vertically integrated system into components to achieve sustainable healthy growth.
- Hard to understand precise mechanisms of every component of entire system.
- Measure and collect big data on inputs and outputs of the system to infer the relationship between them.
- Mathematical tools, e.g., machine learning are available here.
- Key to success is inter-play between mathematics and network engineering.

Thank you for your attention