NSLAB TSINGHUA UNIVERSITY

SMILER
Practical Online Traffic Classification

Presented by *Baohua Yang*

September 14, 2011

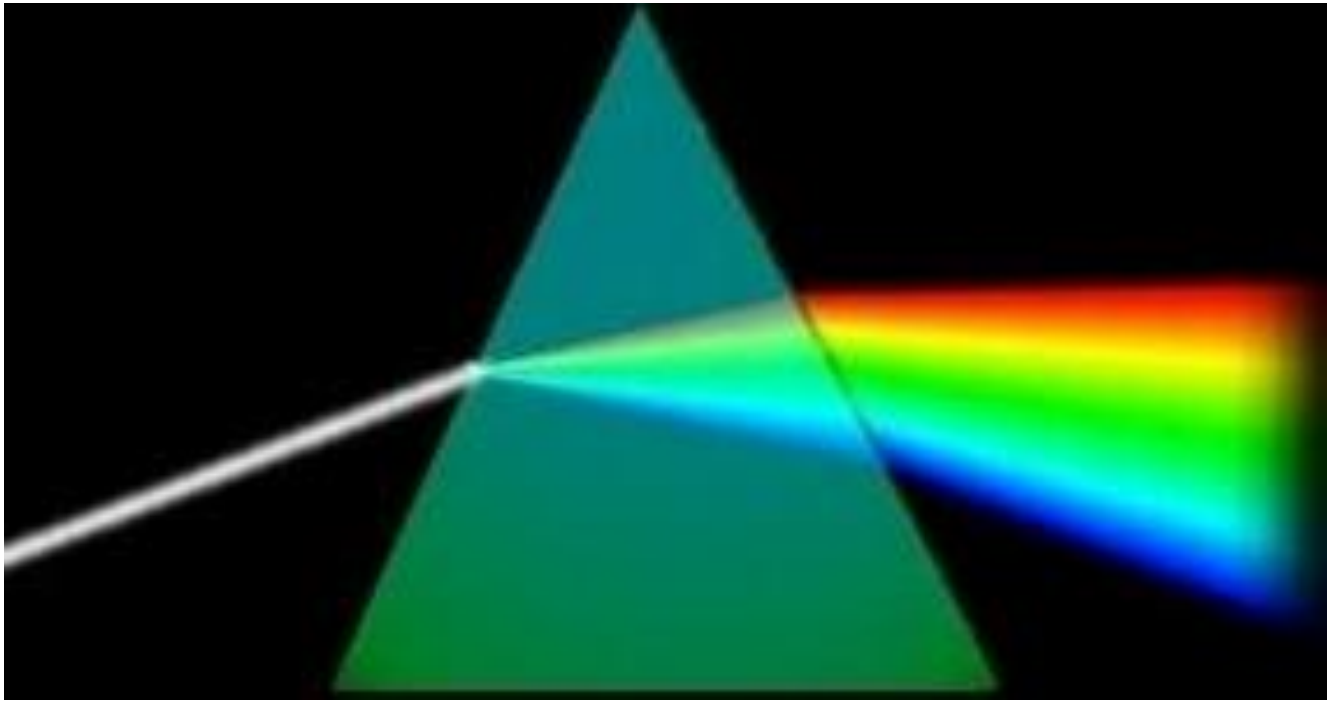
Baohua Yang, Guangdong Hou, Lingyun Ruan, Yibo Xue and Jun Li

Content

- Background
- Theory and Design
- Evaluation Results
- Conclusion

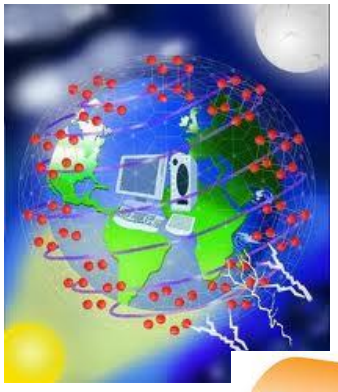
Background

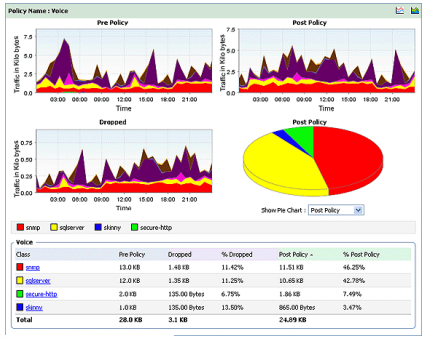
- What is Traffic Classification (TC) ?

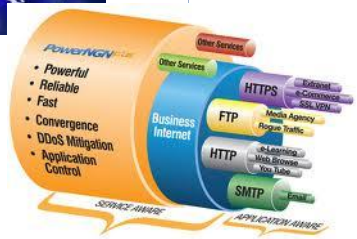


Background

- Why do we need TC?
  - QoS
  - Security
  - Performance
  - And…







Background

- What does a practical TC solution require?
  - Accuracy
  - Early identification
  - Flexibility
  - Speed

Background

	Accuracy	Early Identification	Flexibility	Speed
Port-based	☹	☺	😐	☺
DPI on payload	☺	☹	☹	☹
Traditional Machine learning	😐	😐	😐	☺
SMILER	☺	☺	☺	☺

- SMILER: a SeMi-supervIsed Learning based classifiER to meet all these requirements!

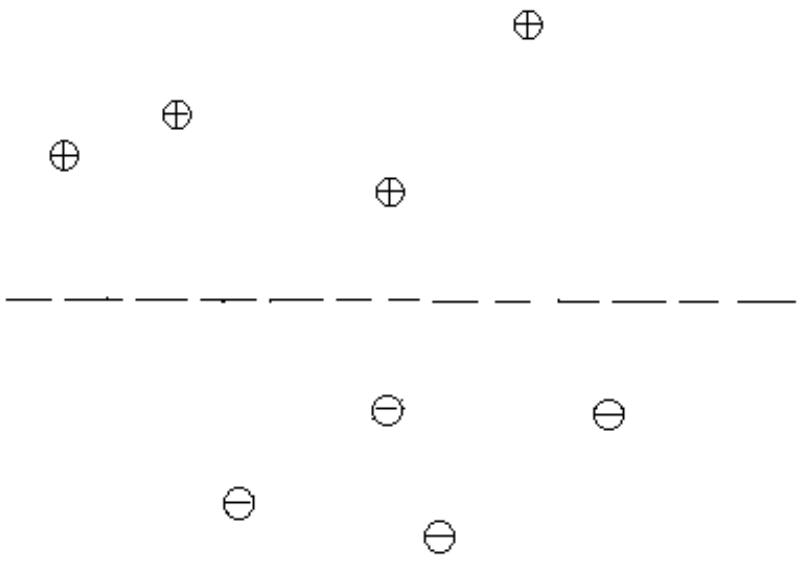
Content

- Background
- Theory and Design
- Evaluation Results
- Conclusion

Why do we choose Semi-supervised Learning?

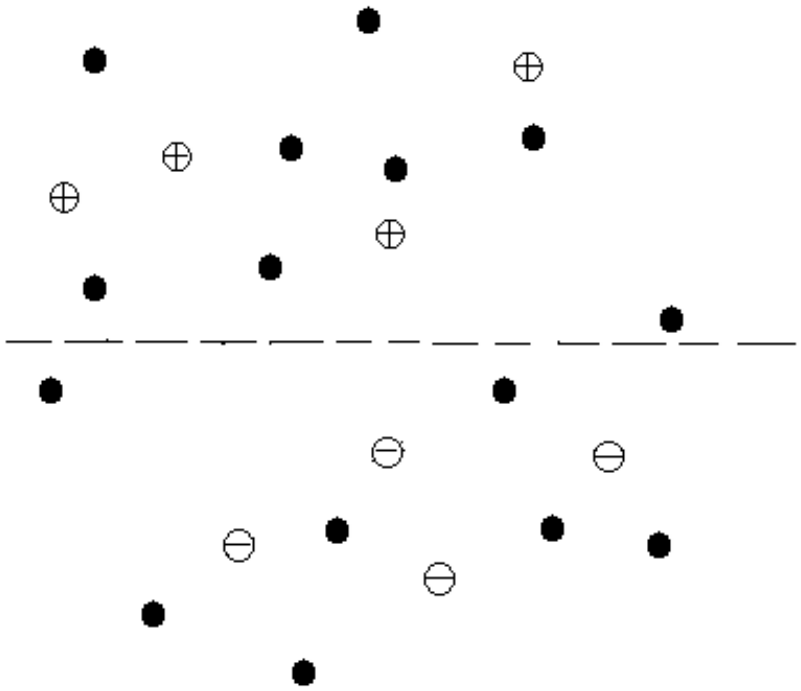
Theory and Design

- Supervised Machine Learning



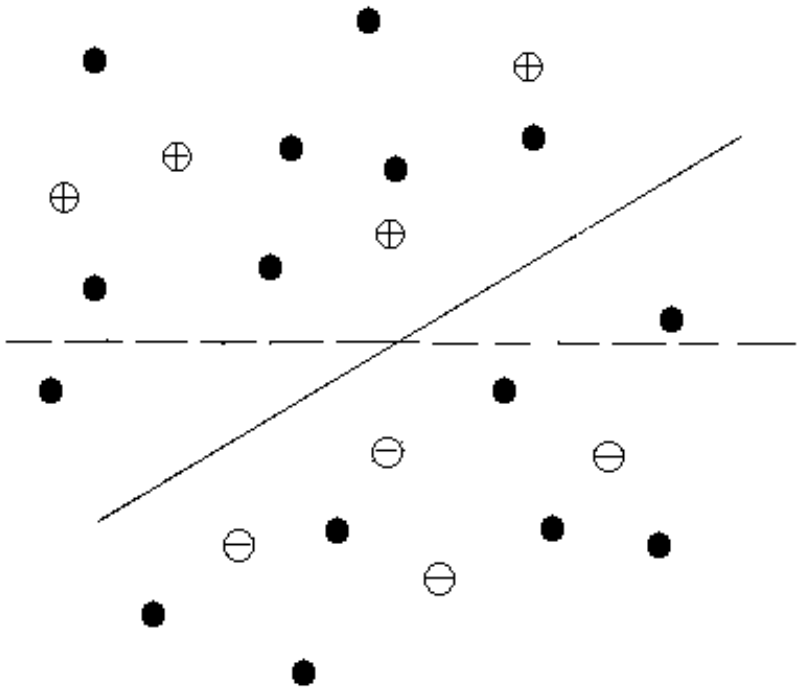
Theory and Design

- Supervised Machine Learning



Theory and Design

- Semi-supervised Machine Learning is better.



Extend 2-class classification to multi-class?

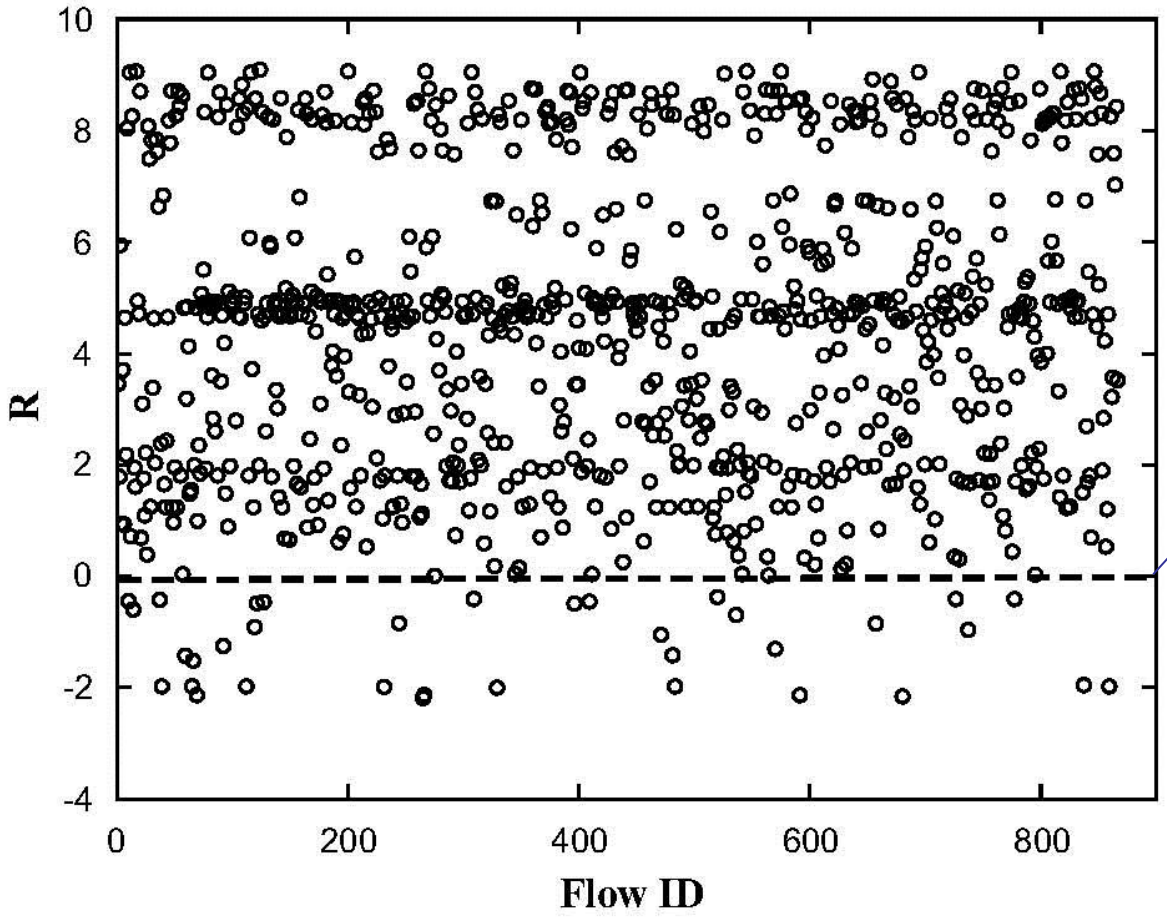
Theory and Design

- From 2-class to Multi-class
  - For each pair in the two categories, a binary-class classifier will be trained. For N classes in total, $N(N-1)/2$ classifiers will be generated.
  - A voting is carried out among all classifiers to gain the final prediction label.

Hybrid Scheme to improve accuracy.

Theory and Design

- Observation



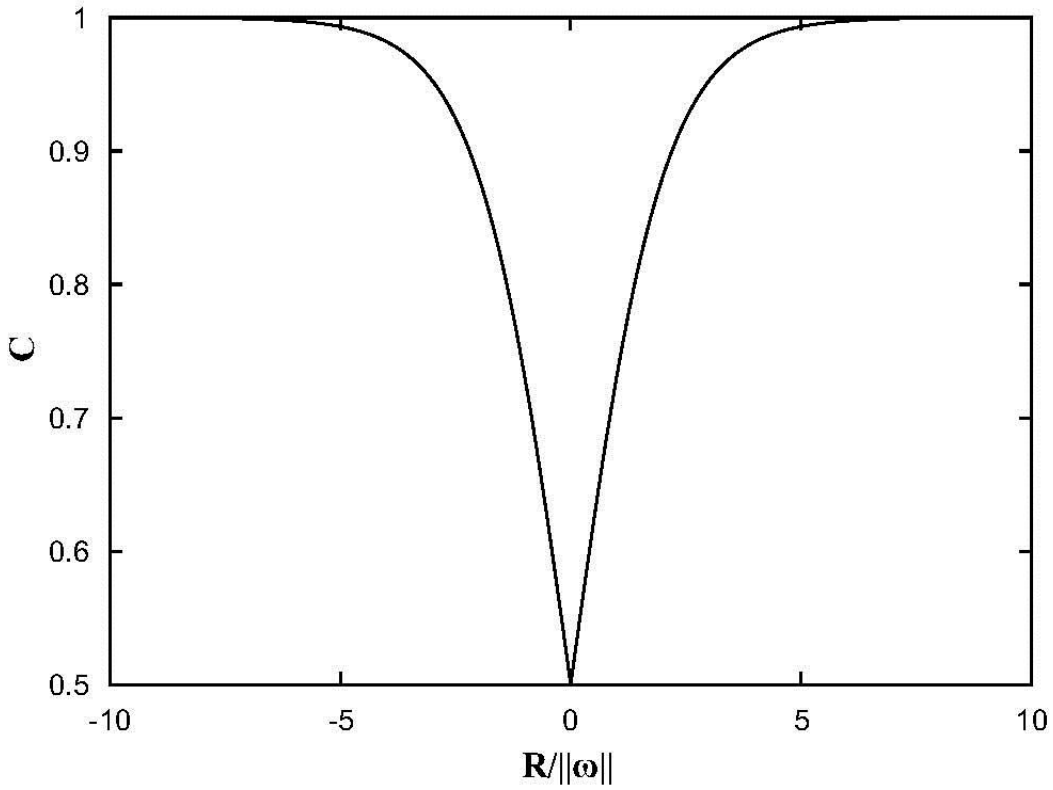
Theory and Design

- *C-factor*
  - A confidence factor by the distance $\frac{|R|}{||W||}$ from the hyper-plane.
  - Range in [0.5,1.0], when $\frac{|R|}{||W||}$ approaches ∞, C should approach 1.0; when $\frac{|R|}{||W||}$ approaches 0, C should be 0.5.

$$C = \frac{1}{1 + e^{-\frac{|R|}{||w||}}}$$

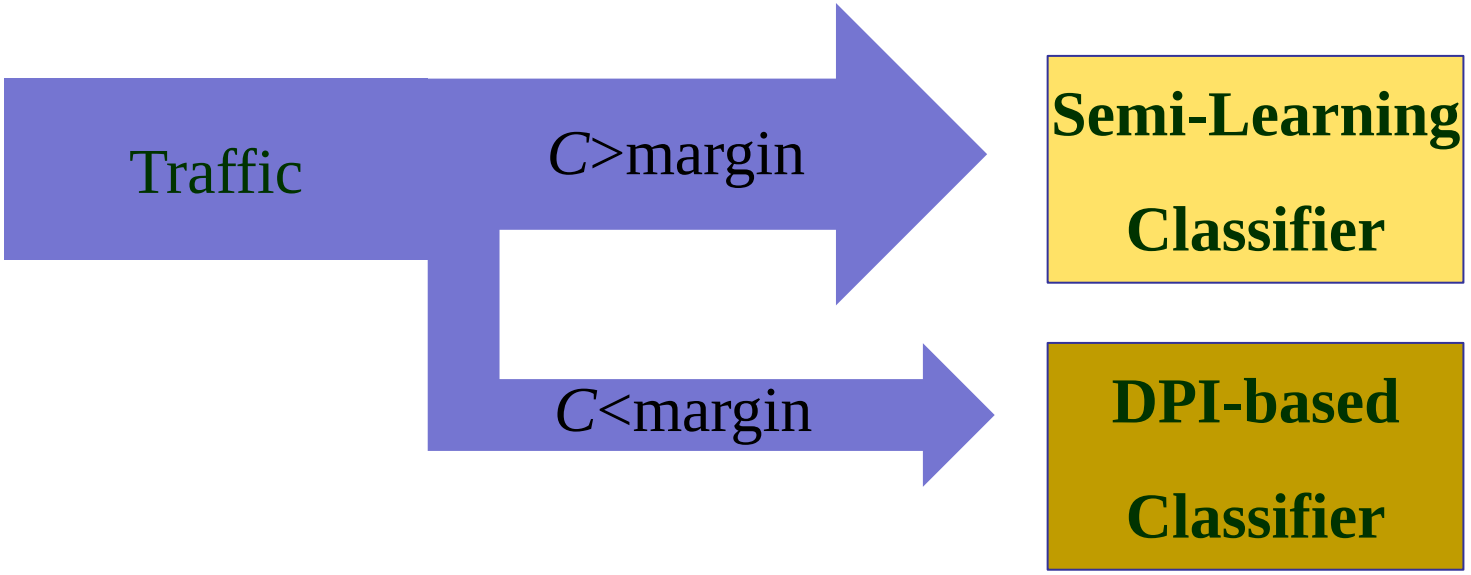
Theory and Design

- *C-factor*



Theory and Design

- Hybrid Scheme
  - Basic Idea: Transfer the traffic with lower *C-factor* value to other accurate classifiers, e.g., DPI-based ones.

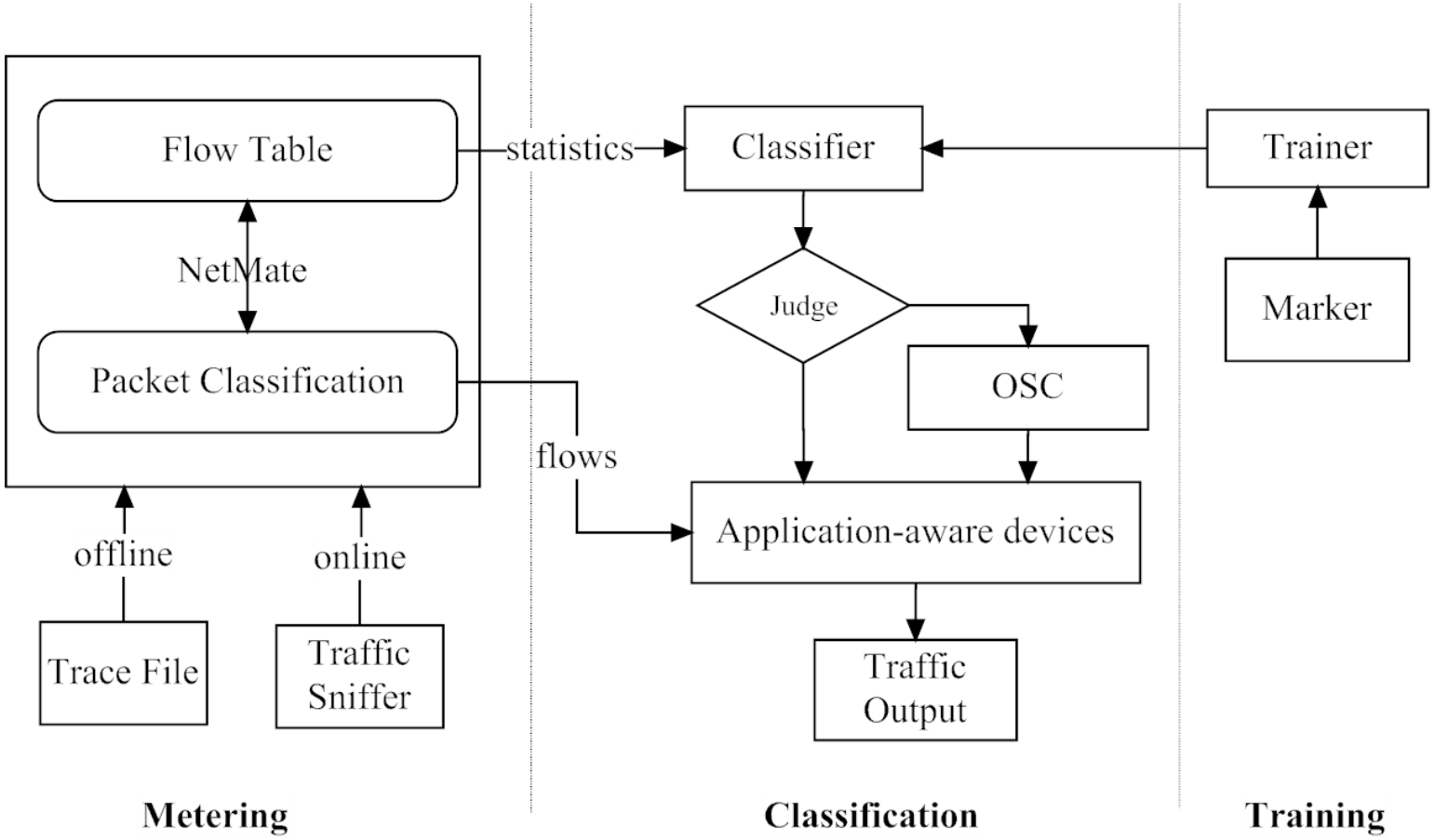


How to handle disordered packets?

Theory and Design

- Packet disorder
  - Disordered packets may happen after multi-path transferring.
  - Most existing techniques are based on packet reassembling, which might increase latency and storage while waiting for unreachable packets.
  - Utilize missing feature classification techniques, e.g., set missing features to 0.

Theory and Design



Content

- Background
- Theory and Design
- Evaluation Results
- Conclusion

Evaluation

- Trace sets
  - Collected on 2008 and 2010 in a large campus network.
  - Classification on 6 representative applications.
    - Web (non-video HTTP)
    - Video (over HTTP)
    - FTP
    - SMTP
    - BT
    - SSH

Evaluation

- Measure parameters
  - E.g., classify traffic of X type and Not-X type

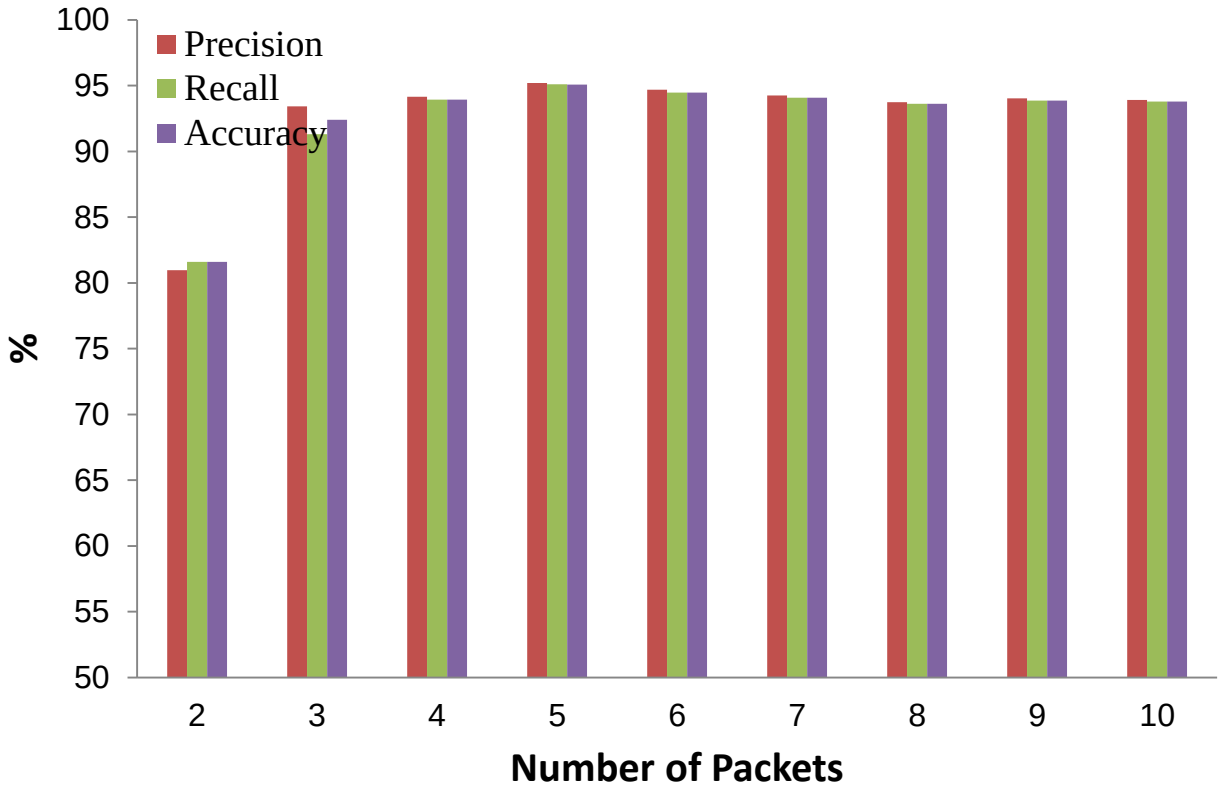
Traffic/Result	**Classify as X**	**Classify as Not-X**
Traffic of X	True Positive	False Negative
Traffic of Not-X	False Positive	True Negative

Evaluation

- Measure parameters
  - Precision
    - TP/(TP+FP)
    - Classified as X, and how much is real X?
  - Recall
    - TP/(TP+FN)
    - Real X, and how much is classified as X?
  - Accuracy
    - (TP+TN)/(TP+TN+FP+FN)
    - Correct classification ratio for all traffic.

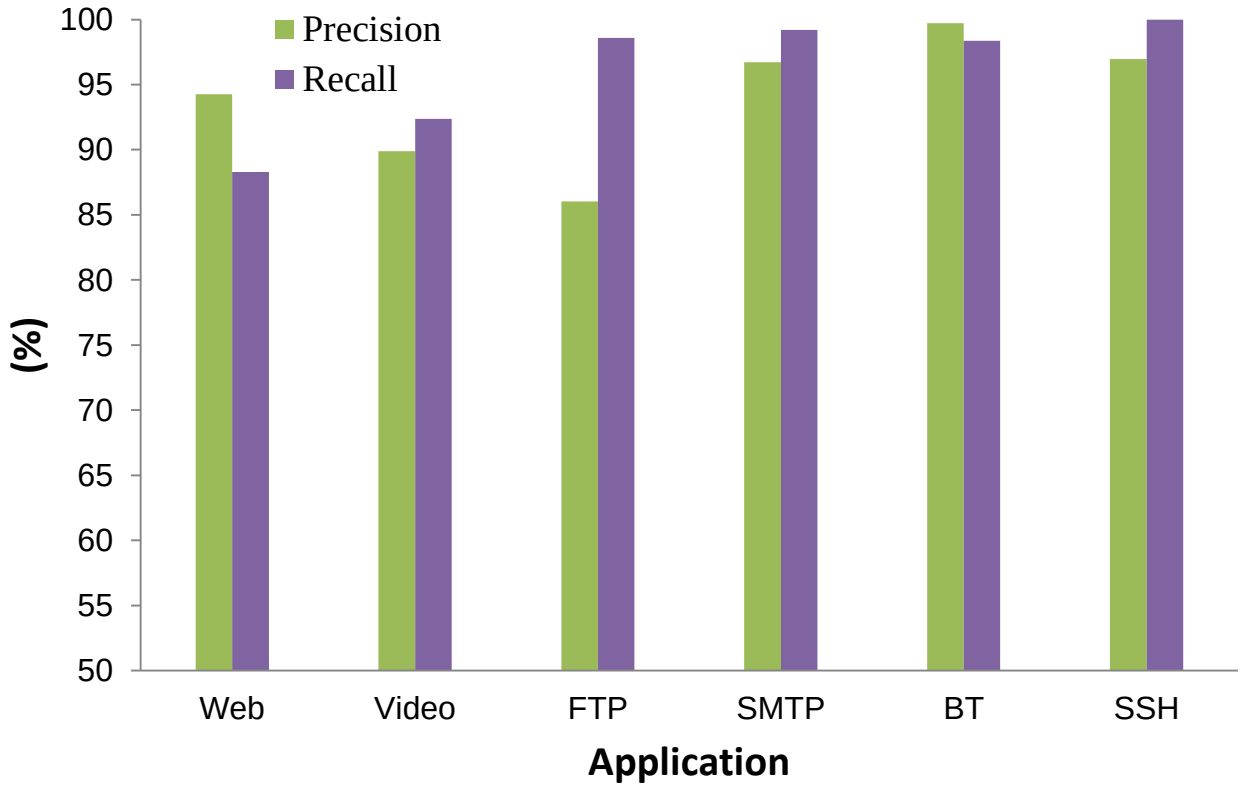
Evaluation

- Results with Different Number of Packets



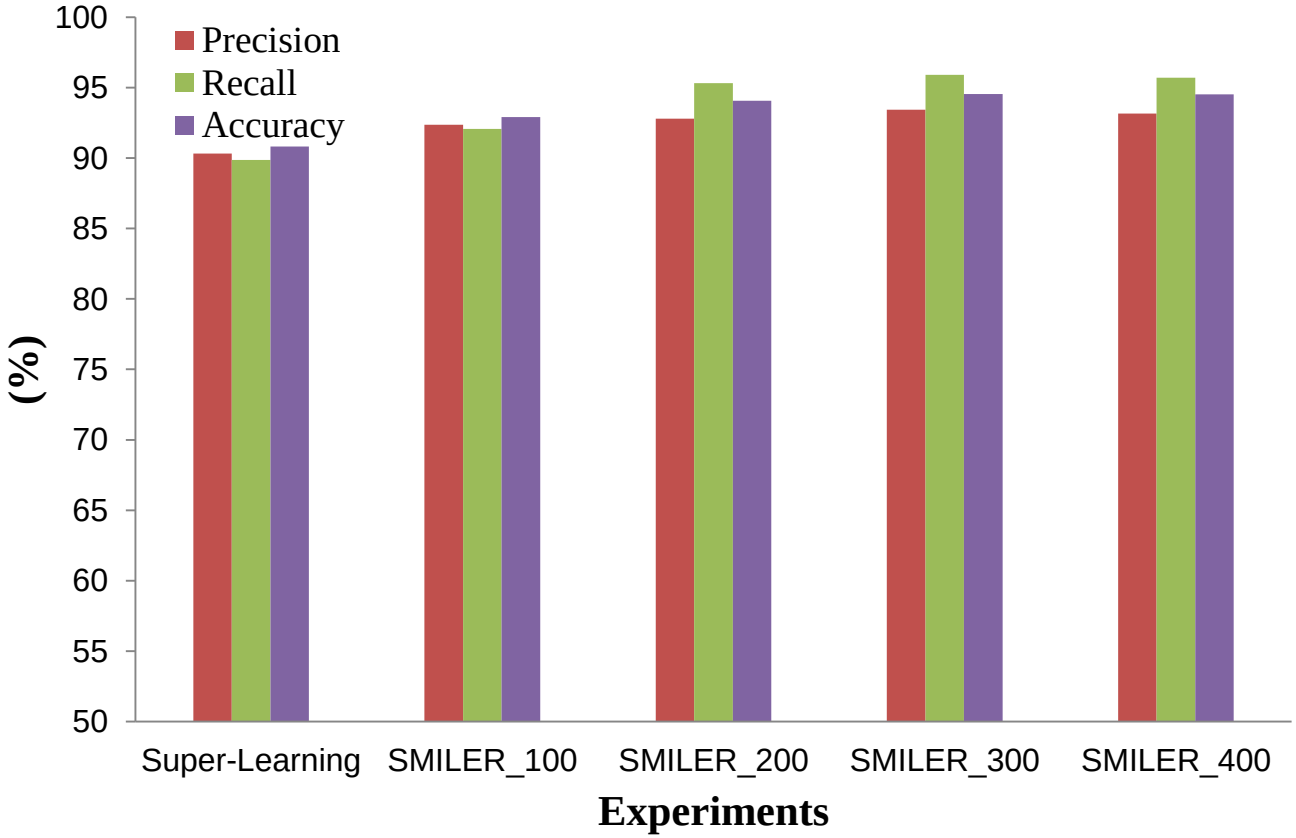
Evaluation

- Results over different applications



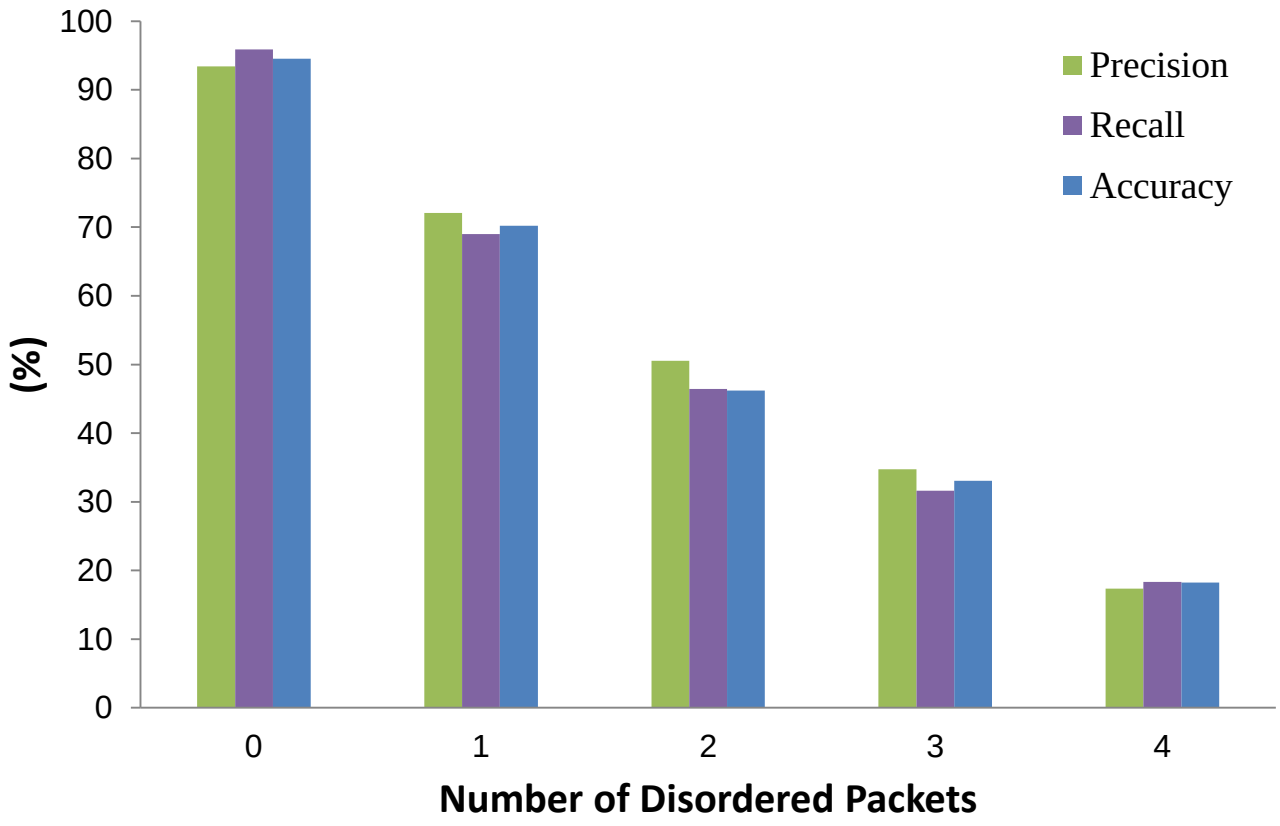
Evaluation

- SMILER vs. Supervised Machine Learning



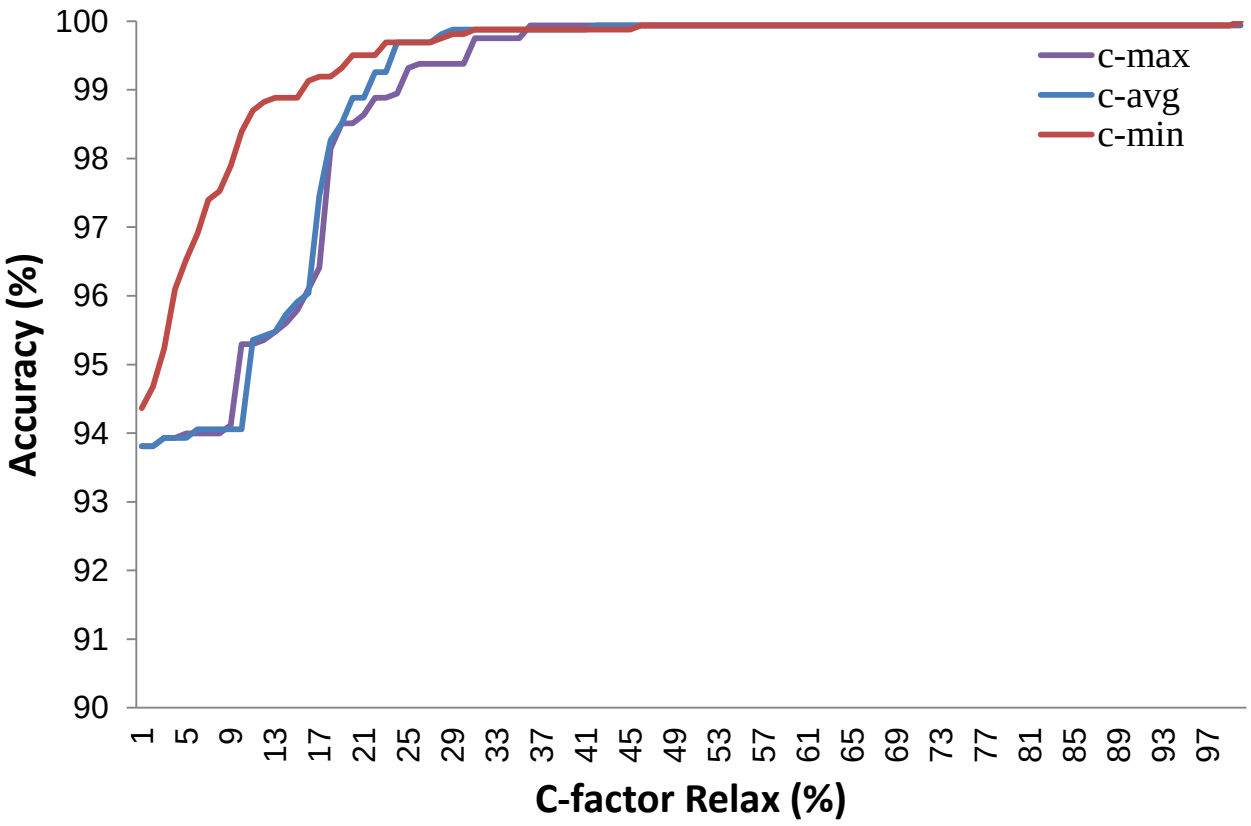
Evaluation

- Disordered Packets



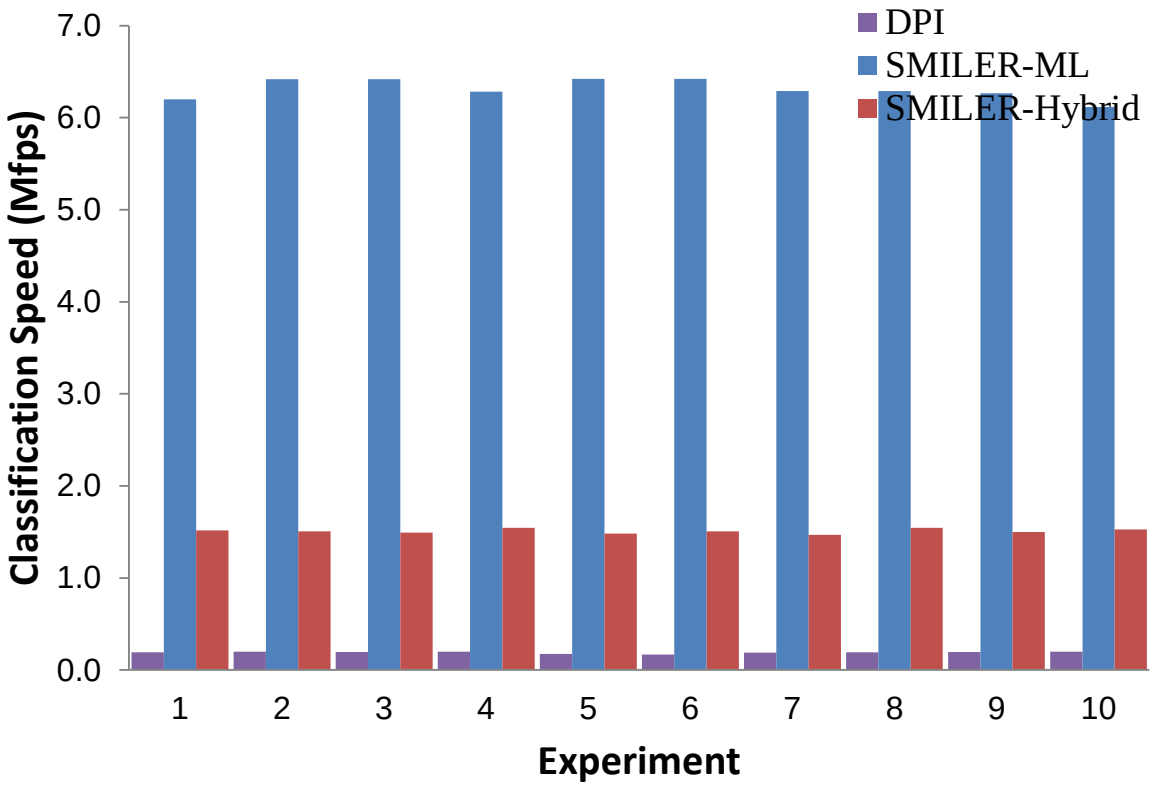
Evaluation

- Hybrid scheme improvement



Evaluation

- Classification Speed



Content

- Background
- Theory and Design
- Evaluation Results
- Conclusion

Conclusion

- SMILER is practical in online classification.
  - Accurate
    - over 95% accuracy with sizes of first 5 packets.
  - Early identification
    - No detection on packet content.
  - Flexibility
    - Easy to integrate with other approaches.
  - Speed
    - 8X~30X over DPI.

Thanks

Any question?