Analysis of Communication Patterns in Network Flows to Discover Application Intent

Presented by:
William H. Turkett, Jr.
Department of Computer Science

WAKE FOREST
UNIVERSITY

FloCon 2013 | January 9, 2013

WAKE FOREST UNIVERSITY

Traditional Traffic Classification Techniques

Port- and payload signature-based classification techniques are increasingly less useful in modern traffic analysis.

Statistical approaches evaluating features such as packet size and interarrival times developed in response.

Traditional HTTP connection:

[src, src prt, dst, dst port, payload]

[10.1.11.58,8754, 10.19.132.45,80,

"GET /index.html"]

HTTP

Modern traffic:

[10.1.11.58,8754, 10.19.132.45,9090,

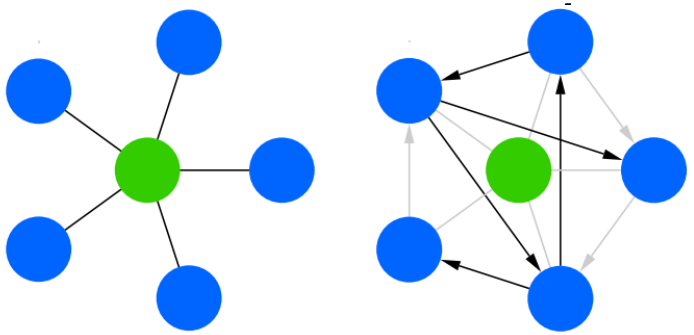
"xZvRmTTlFz"]

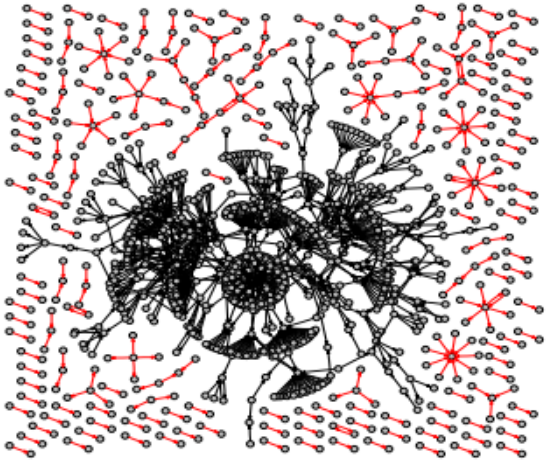
Encrypted payloads

Alternative ports/tunneling

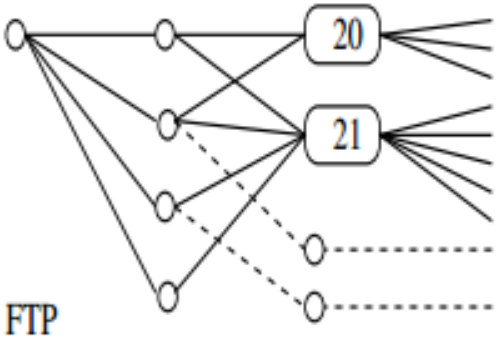
Graph Based Approaches To Traffic Classification

Graph based approaches look at the broader context of interactions (interaction networks instead of topological networks)





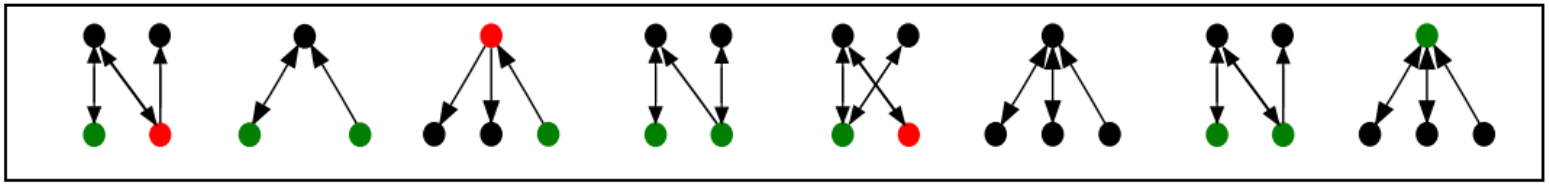
Graption – Traffic Dispersion Graph



BLINC - Graphlet

Karagiannis et al. - BLINC: Multilevel Traffic Classification In The Dark, SIGCOMM Proceedings, 2005.
Iliofotou et al. Graption: Graph-based P2P Traffic Classification At The Internet Backbone, Computer Networks, 2011

Communication Patterns And Motifs

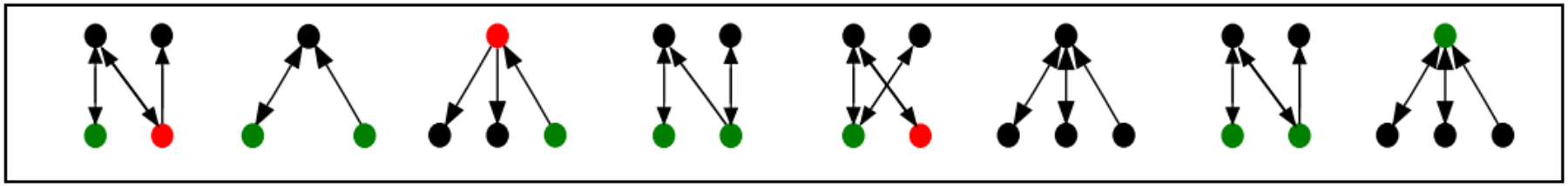


***Motifs* are patterns of interconnections occuring in networks at rates greater than expected by chance.**

Flow-level statistics can be employed to color graph nodes (hosts), allowing for *annotated motifs*

- *Bytes*: {Max, Average, Sum} bytes sent by a host over all connections host involved in
- *Duration*: {Max, Average, Sum} duration of connections host involved in
- *Node Type*: Client, server, or peer activity

Communication Patterns And Motifs



{ 1 0 0 0 1 1 0 0 }

Motif profiles **for a *host* represent in a binary vector which annotated motifs a host participates in**

FANMOD a tool for fast network motif detection

Tools such as *FANMOD* can mine graphs for motifs and determine host-level motif participation

Information Available From Flow Data

The data of interest to build graphs and color nodes is all accessible from flow data:

- Host-host interactions (Src-Dst)
- Summary-level statistics of traffic
  - Number of bytes transferred over connections
  - Duration of connections (timestamps)

SrcIf	SrcIPadd	DstIf	DstIPadd	Protocol	TOS	Flgs	Pkts	Src Port	Src Msk	Src AS	Dst Port	Dst Msk	Dst AS	NextHop	Bytes/ Pkt	Active	Idle
Fa1/0	173.100.21.2	Fa0/0	10.0.227.12	11	80	10	11000	162	/24	5	163	/24	15	10.0.23.2	1528	1745	4
Fa1/0	173.100.3.2	Fa0/0	10.0.227.12	6	40	0	2491	15	/26	196	15	/24	15	10.0.23.2	740	41.5	1
Fa1/0	173.100.20.2	Fa0/0	10.0.227.12	11	80	10	10000	161	/24	180	10	/24	15	10.0.23.2	1428	1145.5	3
Fa1/0	173.100.6.2	Fa0/0	10.0.227.12	6	40	0	2210	19	/30	180	19	/24	15	10.0.23.2	1040	24.5	14

- Assume can capture internal-to-internal and internal-to-external connections

A Deeper Problem: Discovery of Application Intent

Streaming media

Email

HTTP

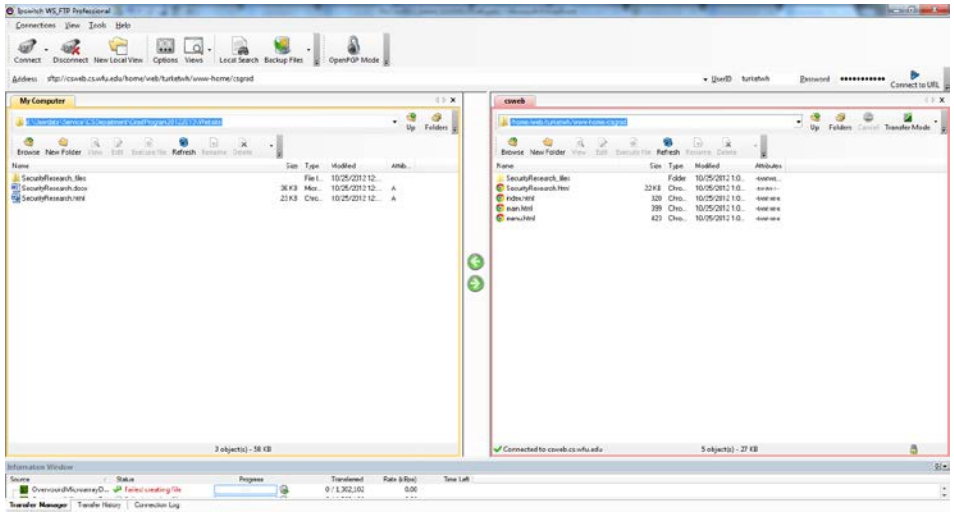
Chat

Browsing

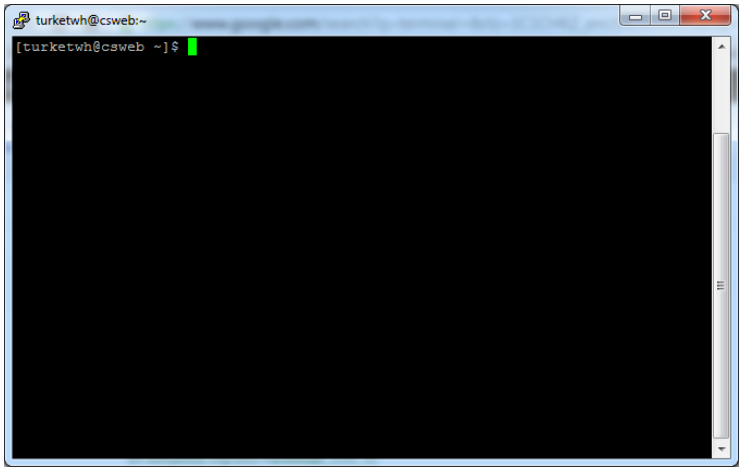
Single network protocols are now commonly employed for a variety of applications (intents)

WAKE FOREST UNIVERSITY

SSH: Application Intent



File Transfer



Terminal

Tunneling

Essence of Approach

Goal is labeling host intent from capture of a window of activity

- Potentially multiple connections within a window of activity
- Assuming that intents are used in *isolation* within a session

As designed currently, prime application is post-mortem analysis of host activity of interest.

Premise of research:

- *Annotated* and *directed* motifs capture significant information about communications
- Hypothesis: Distinct motif usage suggests distinct intent.

Traffic Classification Using Motifs: Initial Work

Our original work in this area (2009) explored separability of individual protocols, not intents.

Modeling approach consisted of:

- Construction of interactions graphs for each protocol
- Node coloring by host type (client/server/peer)
- Host motif profiles were over sets of size three or size four motifs from interaction graphs

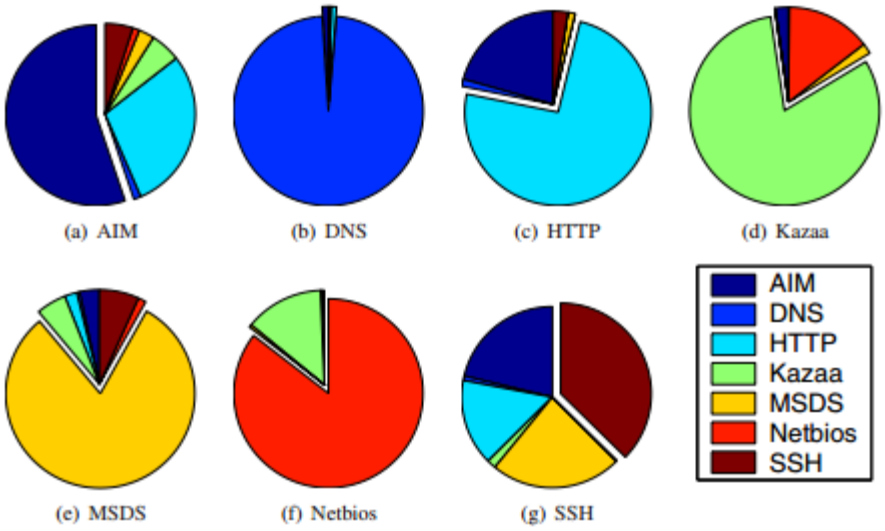
Host-protocol classification approach consisted of:

- Weighted-feature one-nearest-neighbor

WAKE FOREST UNIVERSITY

Protocol Separation Using Motifs

	True AIM	True DNS	True HTTP	True Kazaa	True MSDS	True NetBIOS	True SSH	Precision
Predicted AIM:	**298**	8	56	0	18	0	32	72.33%
Predicted DNS:	7	**632**	3	9	2	0	4	96.19%
Predicted HTTP:	120	14	**676**	0	19	3	23	79.06%
Predicted Kazaa:	5	0	1	**370**	5	34	1	88.94%
Predicted MSDS:	2	4	15	2	**269**	1	1	91.50%
Predicted NetBIOS:	0	1	0	0	2	**700**	0	99.57%
Predicted SSH:	36	0	19	1	57	2	**94**	44.98%
Recall:	63.68%	95.90%	87.97%	96.86%	72.31%	94.59%	60.65%	† **85.70%**



Data Sets For Intent Analysis

Goal is labeling host intent from capture of a window of activity

Properties of publicly available network datasets lead to difficulty in defining gold-standard datasets for training and analysis

- Privacy issues lead to IP shuffling and payload removal

Intent labeling is even harder

WAKE FOREST UNIVERSITY

Experimental Design: Flow Capture

For this work, flows were:

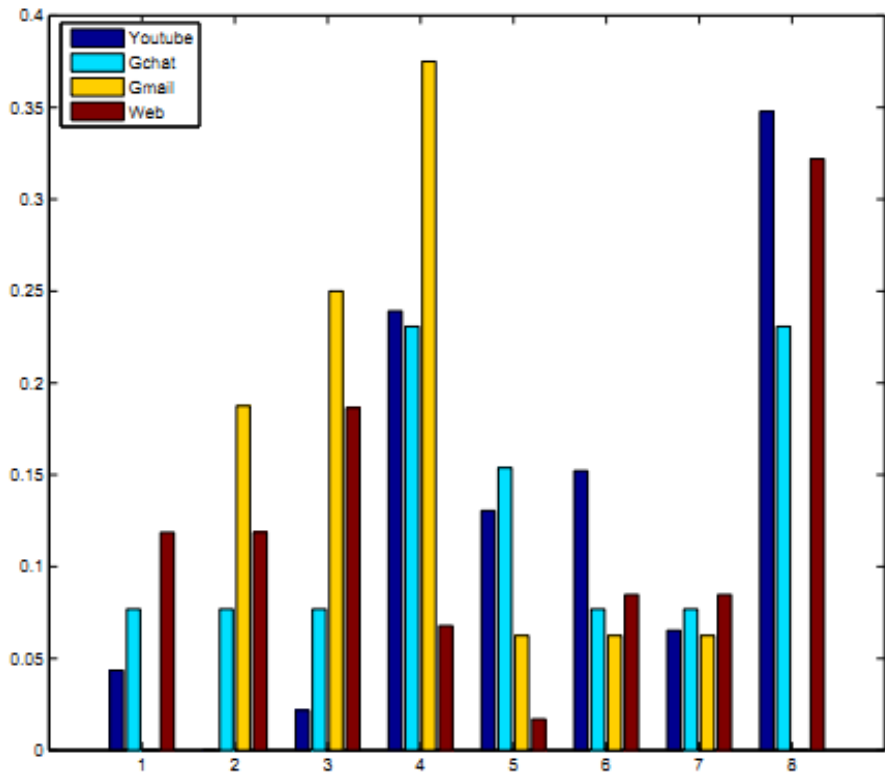
- Collected in-house
- Intents captured in isolation
- Captures automated through AutoIt scripts
- Kept any flows involved in a connection to purported HTTP host (port 80, 8080, 443)

Traffic Type	Source
Streaming media	Youtube
Email	GMail
Chat	GChat
Browsing	Yahoo random link generator

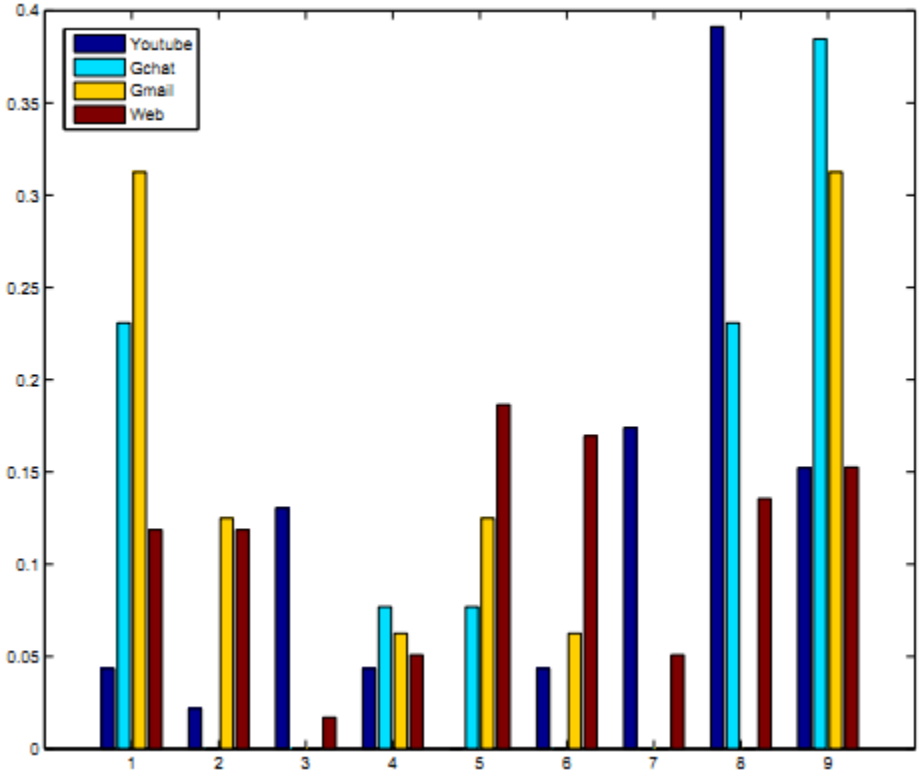
WAKE FOREST UNIVERSITY

Experimental Design: Histograms Of Annotation Statistics

No clear separation of distributions over bytes transferred or connection duration from visualization of flow statistics.



Average Bytes Transferred
(Binned, From Flow Statistics)



Average Flow Duration
(Binned, From Flow Statistics)

Experimental Design: SVM Approach and Results Summary

Support vector machine learning:

- Multiple “one-vs.-all” *support vector machine* models
- Max over model scores
- 10-fold cross validation

Accuracy across flow types (for small sample):

Truth	Total Flows	Node Type Only	Node Bytes + Type	**Node Duration + Type**
Gchat	21	0.71	1.00	**1.00**
Gmail	19	0.00	0.68	**1.00**
Browsing	71	1.00	0.97	**1.00**
Youtube	46	0.00	0.93	**0.94**

Node Duration & Type Results

Confusion matrix for model with best results – the model employing Node Duration and Type:

Label / **Truth**	Gchat	Gmail	Browsing	Youtube
Gchat	21	0	0	0
Gmail	0	19	0	0
Browsing	0	0	71	0
Youtube	3	0	0	43

Conclusions

Building evidence that subgraphs (motifs) of host interaction networks are related to type of activity (intent) being performed by hosts

Flow metrics, traditionally employed by statistical approaches to traffic analysis, can be embedded into graph structures through node coloring

Technology Transfer & Future Work

Online costs of deployment for approach:

- Building the host interaction network from network monitoring over time
- Determination of whether a host is involved in a set of motifs of interest
- Classification model scoring

Next steps:

- Refine traffic generation and collection processes
- Determine lower-limit on data required to accurately reflect a host's activity
- Remove assumption that intents are performed in isolation within a session of activity
- Understand the important motif structures

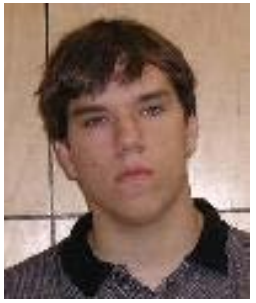
Acknowledgements

Network Security Colleagues at Wake Forest University

Dr. Errin Fulp

Brad McDanel

Lee Bailey



Tim Thomas

National Science Foundation

Grant # CNS-1018191