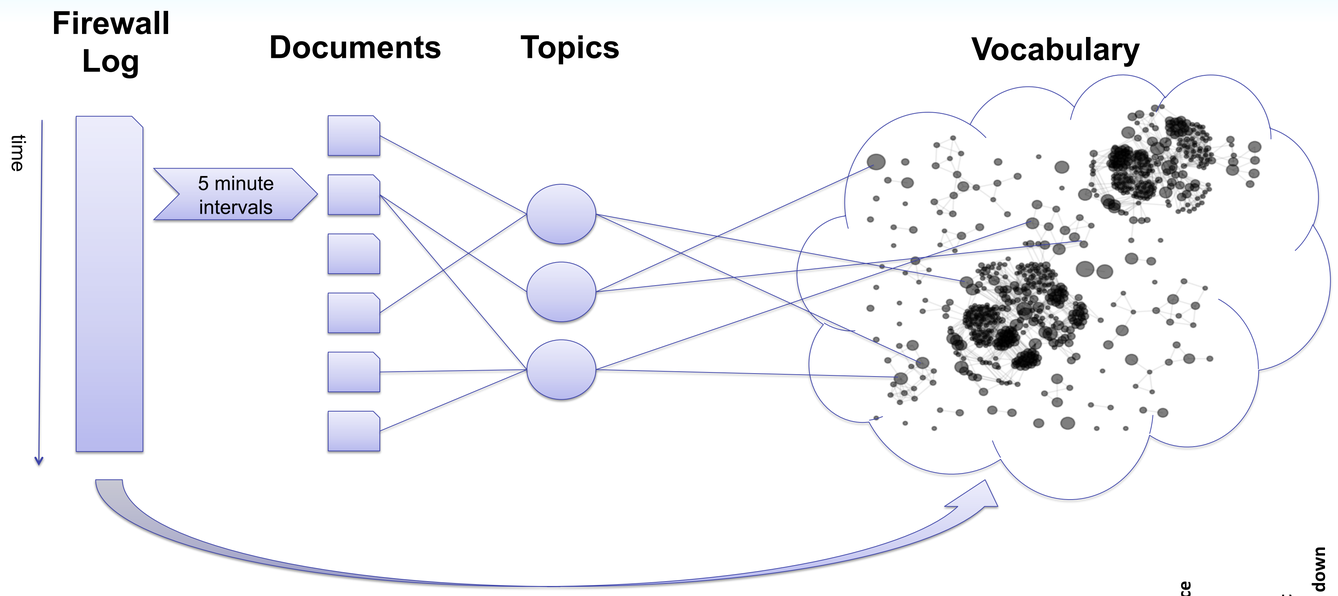
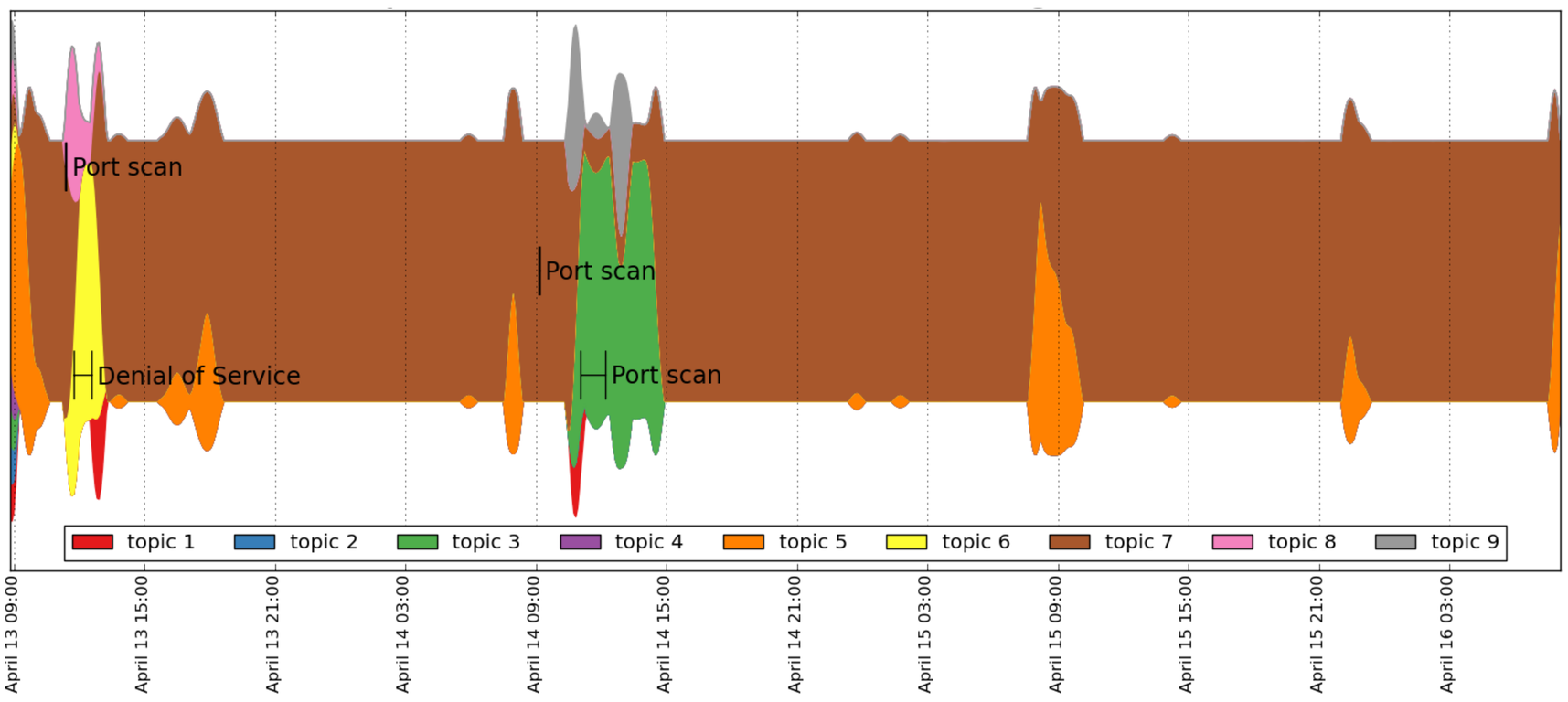
Hot Topics: Information Retrieval for Network Security

Dustin L. Arendt[1,2]
1- Postdoctoral Research Associate of the National Academy of Sciences
2- United States Air Force Research Laboratory, Wright-Patterson AFB, Dayton, Ohio

Method

- **Data Source: VAST 2011 Mini Challenge 2 Firewall Log**
  - ~12M events over 3 days
  - contains multiple attacks (denial of service, port scan)
- **Preprocessing (reduces number of unique log entries)**
  - Port: replaced with # of digits;
  - IP address: replaced with network role
  - Service[:port]: port omitted
  - Message code, source & destination host names omitted
- **Each row in the firewall log is treated as a word**
- **Each 5 minute interval is a document (bag of words abstraction)**
- **Latent Dirichlet Allocation [1,4] used to learn topics**
  - Documents are distributions over topics
  - Topics are distributions over words
  - Words are simplified firewall log entries
- **Topics over time visualized (below) as a stacked graph [2,3]**





Topic	Score	Syslog priority	Operation	Protocol	Source IP	Destination IP	Source port	Destination port	Destination service	Direction	Connections built	Connections torn down
1	0.276	Info	Teardown	TCP	Internet	External Web	4	2	http	(empty)	0	1
1	0.249	Info	Built	TCP	Internet	External Web	4	2	http	inbound	1	0
1	0.051	Info	Built	TCP	Office	Unknown	5	4	tcp	inbound	1	0
2	0.317	Info	Built	TCP	Internet	External Web	4	2	http	inbound	1	0
2	0.281	Info	Teardown	TCP	Internet	External Web	4	2	http	(empty)	0	1
2	0.077	Info	Teardown	TCP	Office	Unknown	5	4	tcp	(empty)	0	1
3	0.303	Info	Built	TCP	Internet	External Web	4	2	http	inbound	1	0
3	0.285	Info	Teardown	TCP	Internet	External Web	4	2	http	(empty)	0	1
3	0.123	Info	Built	TCP	Office	Unknown	5	4	tcp	inbound	1	0
4	0.378	Info	Built	TCP	Internet	External Web	4	2	http	inbound	1	0
4	0.338	Info	Teardown	TCP	Internet	External Web	4	2	http	(empty)	0	1
4	0.084	Info	Built	TCP	Office	Unknown	5	4	tcp	inbound	1	0
5	0.170	Info	Teardown	TCP	Office	Data Center	4	3	epmap	inbound	0	1
5	0.163	Info	Built	TCP	Office	Data Center	4	3	epmap	inbound	1	0
5	0.151	Info	Built	TCP	Office	Data Center	4	5	tcp	inbound	1	0
6	0.467	Info	Built	TCP	Internet	External Web	4	2	http	inbound	1	0
6	0.466	Info	Teardown	TCP	Internet	External Web	4	2	http	(empty)	0	1
6	0.050	Info	Teardown	TCP	Internet	External Web	4	2	http	inbound	0	1
7	0.223	Info	Deny	TCP	Data Center	Office	5	4	tcp	(empty)	0	0
7	0.066	Info	Deny	TCP	Data Center	Office	3	4	tcp	(empty)	0	0
7	0.060	Info	Built	TCP	Office	Unknown	5	4	tcp	inbound	1	0
8	0.521	Info	Built	TCP	Internet	External Web	4	2	http	inbound	1	0
8	0.269	Info	Teardown	TCP	Internet	External Web	4	2	http	(empty)	0	1
8	0.086	Info	Teardown	TCP	Internet	External Web	4	2	http	inbound	0	1
9	0.163	Info	Built	TCP	Office	Data Center	4	3	epmap	inbound	1	0
9	0.155	Info	Teardown	TCP	Office	Data Center	4	3	epmap	inbound	0	1
9	0.095	Info	Built	TCP	Office	Unknown	5	4	tcp	inbound	1	0

Results

- Normal traffic pattern appears as topic 7
- Medium to long duration attacks also apparent
- Additional network events present

Future Work

- Fuse multiple data sources (i.e., Firewall + IDS + PCAP)
- Train topics on normal traffic and/or known attacks
- Interactive D3 version

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:93-1022, 2003.
[2] L. Byron and M. Wattenberg. Stacked graphs-geometry & aesthetics. IEEE Transactions on Visualization and Computer Graphics, 14(6):1245-1252, 2008.
[3] S. Havre, B. Hetzler , and L. Nowell. ThemeRiver: Visualizing theme changes over time. In IEEE Symposium on Information Visualization, pages 115-123, 2000.
[4] D. G. Ribinson. Statistical language analysis for automatic exfiltration event detection. Technical report, Sandia National Laboratories, 2010.