Detecting Malware P2P Traffic Using Network Flow and DNS Analysis

John Jerrim

FloCon 2013

Overview

- More malware using P2P protocols for command and control
- BotTrawler, a research tool for detecting and classifying P2P traffic
- Use of Protocol Transaction Analysis for detection of P2P protocols
- Detection of ZeroAccess and TDLv4 using PTA
- Examination of Zeus using swarm analytics

The Problem: Malware Using P2P

- **Malware toolkits are including P2P as a means to avoid use of DNS for command and control. Examples include:**
  - Zeus v3
  - TDL v4 (Alureon)
  - ZeroAccess
  - Thor (??)
- **We have observed roughly a 10x increase in the number of malware samples using P2P in the past 12 months**

BotTrawler

- **A network flow and analysis research system that fuses multiple data sources including:**
  - YAF for flow creation and payload analysis
  - Associate DNS lookup with flows
  - Reverse DNS &Passive DNS for flows w/o DNS lookups
  - Geo-Location
  - Reputation
    - Public blacklists / spam lists
    - Private blacklists from DNS convictions
  - Binary file analysis
- **Active research project, but some aspects are being weaponized at this time.**

BotTrawler Processing

- **Identify possible P2P flows and group into "P2P sessions"**
- **Create features for classification based on flow, session, and multi-session analysis**
- **Classify vs. known (labeled) P2P applications for both benign and malware P2P**
  - If known, ignore or alert as appropriate
  - If unknown, cluster with other unknowns and test for suspect malware attributes

P2P Classification Feature Attributes

- **Scalable for high speed analysis**
- **No payload analysis (it's encrypted anyway)**
- **Robust Detection – High True Positive, Low False Positive**
- **Make detection avoidance expensive**
  - Require a protocol change rather than a simple port change, for example
- **Use features the enemy cannot easily control or manipulate**
  - Swarm member characteristics are good features
  - Flow rates and periodicity (automation detection) may be useful but are weaker features

Protocol Transaction Analysis (PTA)

- **Based on features created by examining the number of packets and payload exchanged between the local asset and the P2P swarm members via TCP and UDP**
  - Highly repetitive transaction sequences are readily observable with P2P as there are hundreds (or more) connections (think “connection handshakes”)
  - Easily processed and clustered
  - Typically use 3 to 5 unique transaction sequences to identify a P2P application to handle different command/ response sequences in the protocol
  - Some applications require multiple sets of transaction sequences for different behavioral aspects of the application

Connection Related Attributes

- **Connections to external IP addresses**
  - Focus on unique and rare connections
  - Repeated connections to external Ips
  - Avoid use of DNS
- **Swarm analysis**
  - Geographic dispersion
  - Session to session swarm overlap for same asset
  - Swarm overlap with other suspicious or malicious P2P from other assets

Possible Malware Attributes

- **Swarm members often have other malware installed**
  - % of swarm members on spam lists is generally significantly higher than the "noise level" of benign P2P swarms
- **The geographic distribution of swarm members is generally different than benign P2P swarms**
- **Hybrid P2P applications**
  - Hybrid uses a public network for resiliency and a private network as primary C&C
    - Menti (first observed January 2011) appears to be an example of a hybrid P2P: Uses both Tor and P2P

Additional Malware Attributes

- **Contextually associate P2P traffic with other malware behavior associated with the asset:**
  - P2P traffic begins shortly after (often within seconds) of a suspicious file download
  - Other suspicious activity may also be noted starting near or after the compromise (differential asset behavior):
    - Spamming
    - ClickFraud Activity
    - DoS participation

P2P Classifications

- **General Purpose P2P**
  - BitTorrent
  - eMule
  - Tribbler
  - And many others…

P2P Classifications

- **Specific Purpose P2P**
  - Benign or commercial
    - Skype
    - Spotify
    - And many others
  - Malware
    - ZeroAccess
    - Zeus v3
    - TDL v4
    - And a few others

Specific Purpose P2P

- Are often easily identified by DNS, reverse DNS or passive DNS means as they generally do not try to hide – unless they are malicious
- Swarms are often small (< 100) with some or significant overlap of swarm members between P2P sessions
- Swarms may be highly localized. For example, Spotify uses minimal distance algorithms to reduce propagation delays

General Observations

- **All members of a malware P2P swarm have been compromised with the same malware**
  - Detect one and you will quickly identify hundreds up to tens of thousands of compromised assets
- **P2P Protocols are reused by malware operators. TDLv4 uses the identical P2P protocol as ZeroAccess**
  - Identifying the technology and may identify the primary operator behind the malware, but may not identify the exact compromise

ZeroAccess

- **A rapidly growing click-fraud botnet that uses significant user bandwidth**
  - Over 2 million nodes estimated world-wide in November, 2012
  - Makes extensive use of P2P
  - Appears to be closely related to TDL v4 as it uses the same P2P protocol

Detecting TDLv4 and ZeroAccess

- **Using PTA as primary detection mechanism**
  - Created transaction sequence sets for three variants of the protocol as "labeled data" for the test
  - Simple decision tree for detection:
    - Sequences must be in the "top 5" for the P2P session
    - Three or more unique transaction sequences must be observed
    - Of the three, two must be bidirectional transaction sequences
    - Rank ordered detection is preferred for high confidence

PTA Detection Test Results

- **182,097,625 P2P flows clustered into 132,015 P2P Sessions over a six day period**
  - 168,188 flows in 86 P2P sessions on 49 assets were identified as malware using P2P. All 49 assets were confirmed as infected by the customer (100% True Positive)
  - Transaction Sequence Statistics:
    - An average of 1955 labeled transaction sequences were observed for the P2P sessions classified as malware
    - An average of 1188 labeled bidirectional transaction sequences observed per malware P2P session
    - Only 909 labeled transaction sequences were observed in the remaining 131,992 P2P sessions – all unidirectional
    - There were zero(!) labeled bidirectional transactions observed in the 131,992 non-malware P2P sessions

Zeus v3 BotNet

- Zeus is a botnet focused on banking and financial theft. Use of P2P started early in 2012 when v3 was released.
- Provides a good example of repeated swarm membership for a period of time. Identical swarms have not been observed on benign P2P applications.
- There is a strong indicator of a download containing a list of new swarm members followed by changes in subsequent swarms
- Swarm members exhibited significantly higher spam list rates than background noise.

DAMBALLA
Take Back Command-and-Control

Zeus Multi-Session Swarm Statistics

Session Start	LastTime	IntPkts	IntPayload	ExtPkts	ExtPayload	New IPs	Total
3/15/12 18:34	3/15/12 18:39	950	23912	905	12366	28	31
3/15/12 18:56	3/15/12 19:09	920	17310	901	8020	1	33
3/15/12 19:25	3/15/12 19:39	944	23532	871	8758	1	33
3/15/12 19:55	3/15/12 20:14	1623	26570	1570	8436	0	33
3/15/12 20:30	3/15/12 20:44	1022	36858	1213	136488	9	37
3/15/12 21:07	3/15/12 21:19	890	23240	829	7778	0	29
3/15/12 21:35	3/15/12 21:54	1780	26268	1744	8412	0	31
3/15/12 22:12	3/15/12 22:24	896	15542	888	9032	0	27
3/15/12 22:40	3/15/12 22:59	1724	29314	1648	7962	0	30
3/15/12 23:15	3/15/12 23:29	900	16298	867	6924	0	25
3/15/12 23:45	3/16/12 0:09	2762	72408	2884	162204	37	73
3/16/12 0:26	3/16/12 0:44	1812	35898	1726	9186	0	38
3/16/12 1:00	3/16/12 1:19	1820	29488	1966	102296	0	38
3/16/12 1:37	3/16/12 1:54	1744	27976	1665	8660	0	37
3/16/12 2:10	3/16/12 2:24	951	21502	898	7180	0	29
3/16/12 2:46	3/16/12 2:59	888	17254	1043	82294	0	26
3/16/12 3:16	3/16/12 3:29	966	31184	1128	117210	7	33
3/16/12 3:50	3/16/12 4:04	932	21334	1059	86596	0	28

Key Results

- Identifying new P2P malware works best when intelligently fusing data from a broad range of data sources including network flow and derived features, DNS, binary analysis, swarm analysis, differential behavioral analysis, and reputation systems.
- PTA shows great promise for extracting new information from network flow data to aid in malware and application detection.
- Multi-session swarm analysis provides additional insight into how the botnet is being utilized.

Questions?

?

john.jerrim@damballa.com or on LinkedIn