Alternate Name:

Welcome to the Curse of Dimensionality

By Dennis

Phase 0:
Creating List of Features

Creating List of Features

Conversation

- Packet Info
  - All Information Contained in Packets
- Packet Characteristics
  - Bytes in per packet
  - Bytes out per packet
  - Randomness of bytes in per packet
  - Randomness of bytes out per packet
  - Number of packets per session
- Session Level
  - Randomness of number of packets per session
  - Bytes in per session
  - Bytes out per session
  - Randomness of bytes in per session
  - Randomness of bytes out per session
  - Duration of session
  - Pause time
- Certificate (http://www.netresec.com/?page=Blog&month=2011-07&post=How-to-detect-reverse_https-backdoors)
  - openssl x509 -inform DER -noout -text -in

Flow Data

- Number
  - Conversations per domain
- Protocol
  - Uniqueness
  - Protocol usage
- Time
  - Requests per minute/hour/day
  - Avg. time between requests to a domain
  - Randomness of time between requests
    - Detectable periodicity
  - Time of Day
  - Spike Detection
- Size
  - Avg size
  - Bytes in per minute/hour/day
  - Bytes out per minute/hour/day

Content

- Perceptual hashing
- Longest common substring
- N-grams
- Entropy check
  - Encrypted?
  - Encoded?
  - Steganography?

Domain Name

- Length of address
- Length of subdomains
- Number of subdomains
- Domain generation algorithms
  - Number of unique directories/pagename per domain
  - Frequency of occurrence of directories/pagename length
  - Number of directories/pagename per root domain
  - FQDNs per domain
- Encrypted
- Encoding
  - Entropy of subdomain
  - Wavelet shenanigans
  - Character frequency analysis
  - Percentage of numerical characters in domain name
  - Encoding Detection
    - Base64?
    - Base32?
    - 5-bit?

Domain Reputation and Owner Fingerprinting

- Alexa Rank
- Asn-query
  - Country
  - Registration Date
  - Registrar
  - Owner
- Whois
  - REGISTRANT INFORMATION
  - REGISTRATION INFORMATION
  - ADMIN CONTACT
  - TECHNICAL CONTACT
  - ORGANIZATION NAME
  - NAME SERVERS
- Reputation
  - Actual Domain
  - Hosting provider
- Geo-IP

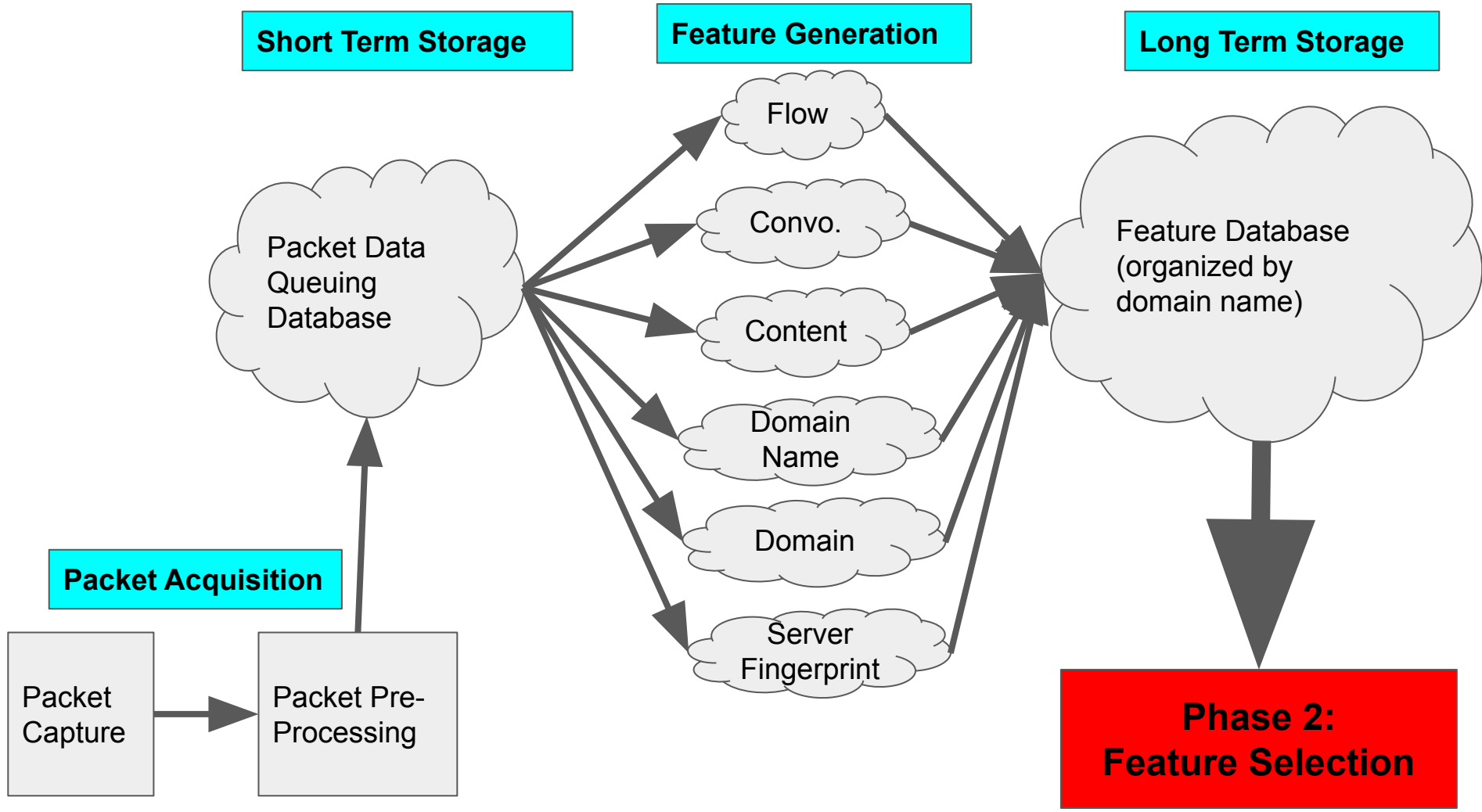
Server Fingerprinting

- Manual request attempts
  - Check responses
- httprint
- httprecon
- nmap Scans
  - https://nmap.org/nsedoc/categories/discovery.html
- **FuzzyFingers**
  - Fuzzing C2 Domains For Unique Outputs
- **Project Omarax**
  - Visit C2 Domain, screenshot, analyze

Phase 1:
Feature Generation

Phase 1: Feature Generation

During this phase the vector created in phase 1 is evaluated and redundant and irrelevant features are discarded. Feature selection has many benefits including: improving the performance of learning modules by reducing the number of computations and as a result the learning speed; enhancing generalization capability; improving the interpretability of a model, etc. Feature selection can be done using a wrapper approach or a correlation-based filter approach. Typically, the filter approach is faster than the wrapper approach and is used when many features exist. The filter approach uses a measure to quantify the correlation of each feature, or a combination of features, to a class. The overall expected contribution to the classification is calculated and selection is done according to the highest value. The feature selection measure can be calculated using many techniques, such as gain ratio (GR); information-gain (IG); Fisher score ranking technique and hierarchical feature selection .


Short Term Storage
Feature Generation
Long Term Storage
Packet Data Queuing Database
Flow
Convo.
Content
Domain Name
Domain
Server Fingerprint
Feature Database (organized by domain name)
Packet Acquisition
Packet Capture
Packet Pre-Processing
Phase 2: Feature Selection

Conversation

- Packet Dissector
- Certificate Dissector
  - Certificate Validation Path
    - Parent, parent’s parent
    - Reputation

Flow Data

- Flow Dissector
  - Number
- Protocol
  - Uniqueness
  - Protocol usage
- Time Dissector
  - Requests per minute/hour/day
  - Avg. time between requests to a domain
  - Randomness of time between requests
    - Detectable periodicity
  - Time of Day
    -

Content

- Pull out data
- Run checks

Domain Name

- Pull out domain names, subdomains
  - Length of address
  - Length of subdomains
  - Domain generation algorithm checks
  - Encoding checks

Domain Reputation and Owner Fingerprinting

- Alexa Rank api
- Asn-query and parse
- Whois and parse
- Reputation api
- Geo-IP and parse

Server Fingerprinting

- Manual request attempts
  - Check responses
- httprint and parse
- httprecon and parse
- nmap Scans and parse
- **Project Omarax**

Phase 2: Feature Selection

Phase 2: Feature Selection

http://machinelearningmastery.com/an-introduction-to-feature-selection/
http://jmlr.csail.mit.edu/papers/volume3/guyon03a/guyon03a.pdf

During this phase the vector created in phase 1 is evaluated and redundant and irrelevant features are discarded. Feature selection has many benefits including: improving the performance of learning modules by reducing the number of computations and as a result the learning speed; enhancing generalization capability; improving the interpretability of a model, etc. Feature selection can be done using a wrapper approach or a correlation-based filter approach. Typically, the filter approach is faster than the wrapper approach and is used when many features exist. The filter approach uses a measure to quantify the correlation of each feature, or a combination of features, to a class. The overall expected contribution to the classification is calculated and selection is done according to the highest value. The feature selection measure can be calculated using many techniques, such as gain ratio (GR); information-gain (IG); Fisher score ranking technique and hierarchical feature selection .

Phase 3: Creating a Classifier

Phase 3: Creating a Classifier

The last phase is creating a classifier using the reduced features vector created in phase 2 and a classification technique. Among the many classification techniques, most of which have been implemented in the Weka platform, the following have been used: artificial neural networks (ANNs), decision tree (DT) learners, nave-Bayes (NB) classifiers, Bayesian networks (BN), support vector machines (SVMs), k-nearest neighbor (KNN), voting feature intervals (VFI) , random forest