Infoblox
NEXT LEVEL NETWORKING

“Looking Ma, No Malware”

Unsupervised Learning Methods to Understand Certain DDoS Attacks

Speaker: Renée Burton

FloCon 2020

Classifying Malicious Actors: classic problem

In intelligence analysis, we are frequently motivated to understand the actors and malware behind malicious behaviors. This can lead to better detection and threat analysis. We look to:

- Understand the landscape of the attacks
- Look for ways to identify similar malware
- Identify similarities across attacks and malware
- Detect or assign a label to malware
- Reverse engineer
- Assign attribution

Classifying Malicious Actors: Data Science

- Understand the landscape of the attacks
  - Statistical analysis of metadata around attacks
- Look for ways to identify similar malware
  - Exploratory data analysis and feature engineering
- Identify similarities across attacks and malware
  - Clustering
- Detect or assign a label to malware
  - Classification (binary and multi-class)
- Reverse Engineer
- *Creating toolboxes to use against different problem sets*

Generally harder

Sharing the Toolbox

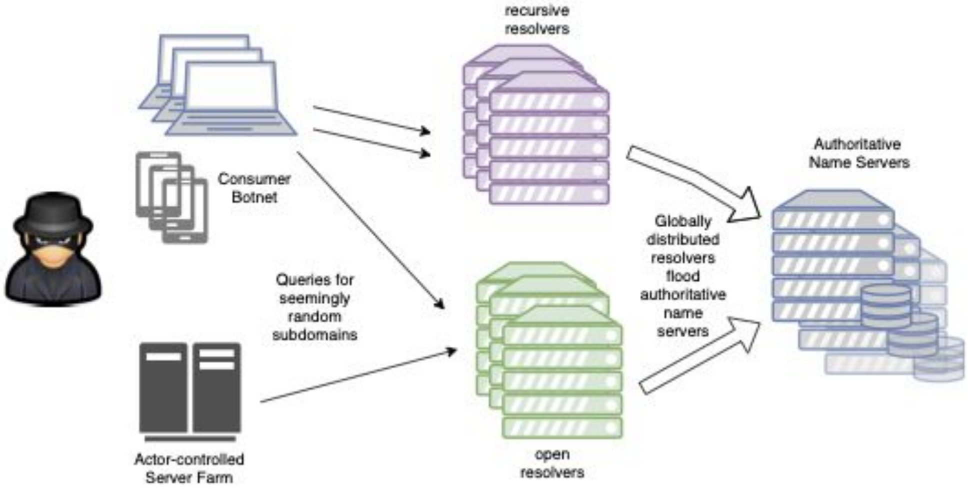
- Frequently we don’t have any labeled data (no malware) ….
- But we might have a lot of observations…
- Unsupervised learning techniques can help us understand
  - Actors
  - Actor Techniques, Tactics, and Procedures

This talk shares some of our tools in the context of

DNS-Based DDoS Attacks

Random Subdomain* DDoS Attacks

- Use “random” subdomains to overload authoritative name servers
  - Resource exhaustion attack
- Attacks against authoritative name servers; domain doesn’t matter really


recursive resolvers
Consumer Botnet
Authoritative Name Servers
Globally distributed resolvers flood authoritative name servers
Queries for seemingly random subdomains
Actor-controlled Server Farm
open resolvers

Goal: Identify and label attacks

We postulate that there are N uniquely identifiable software systems being used in these attacks.

Queries look like:

random_prefix.attack_domain

or

random_prefix.*fixed_label*.attack_domain

* also referred to as Slow Drip or Water Torture DDos attacks

Sample Attack Queries from Various Domains

0k3.hfax.com.	5
t6q.hfax.com.	2
questioned.bjbgp.bjbgp.hfax.com.	2
cxj.bjbgp.hfax.com.	2
lzd.bjbgp.hfax.com.	2
fao.bjbgp.hfax.com.	2
9gi.hfax.com.	1
171j.bjbgp.hfax.com.	1
2ock.bjbgp.hfax.com.	1
krv.bjbgp.bjbgp.bjbgp.hfax.com.	1

us85.91y.com.	10
a34t.bjbgp.91y.com.	7
7c82.91y.com.	3
cp7.bjbgp.91y.com.	2
01uq.bjbgp.91y.com.	2
uh5t.bjbgp.91y.com.	2
dwu9.91y.com.	2
g1we.91y.com.	1

Queries USED TO look like:

random_prefix.attack_domain

or

random_prefix.fixed_label.attack_domain

qiv.wan.douyu.com.	3
worst.bjbgp.g.wan.douyu.com.	3
2xqy.passport.douyu.com.	3

i92.skeeball-arcade.scopely.com.	1
china.scopely.com.	1
books.scopely.com.	1
deal.yahtzee-with-buddies.scopely.com.	1
songs.tech.scopely.com.	1
gw2.wheel-of-fortune.scopely.com.	1
ab4.scopely.com.	1

- Prefix isn't random; can contain dictionary words.
- Fixed label is no longer fixed. Can vary.

The problem

- take ~800M records like the ones on previous slide
- put them into buckets of attacks generated the same way
- where an attack is (date, domain)

The solution

- use machine learning: “unsupervised learning” technique can cluster, or group, attacks that have similar “features”

Slow Drip Feature Engineering

- Problem is really hard -- we only have small observables, no malware
- Approaches we can take include:
  - Landscape statistics, e.g., popularity and co-occurrence of attacked domains, timing, etc.
  - Qname statistics, e.g., how long, how diverse, number of labels, etc.
  - Similarity statistics, e.g, overlaps in attacks and prefixes
  - Time series analysis
  - …..

“good” clustering requires many features.

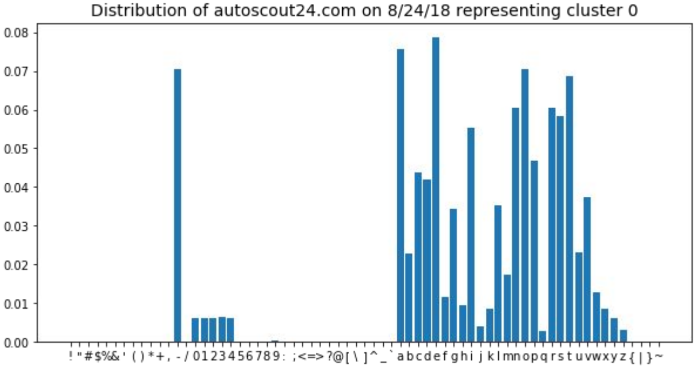
Character Distributions as a Feature

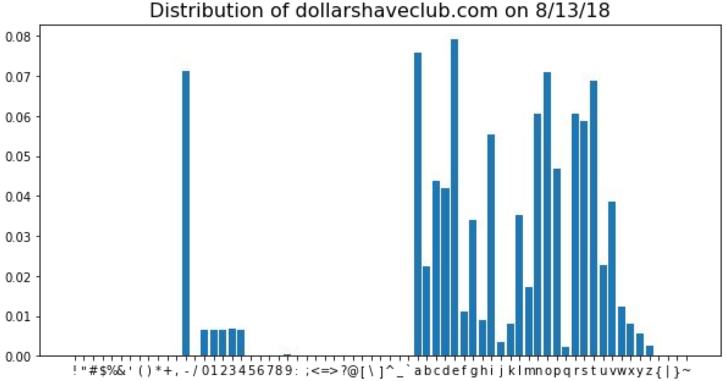
One natural feature to analyze is the characters of the random prefix.

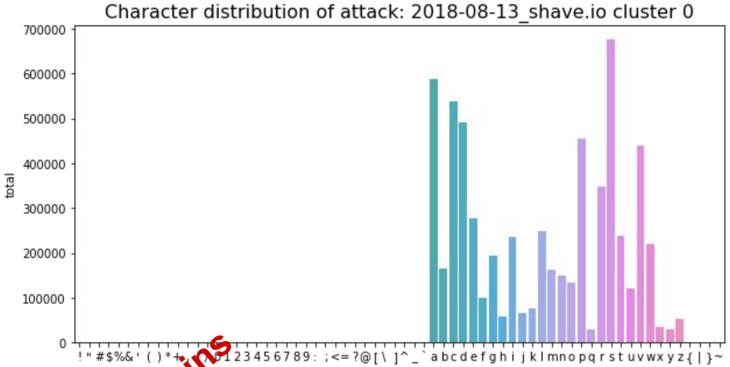
- Common approaches measure digit ratios, entropy, etc.
- Here we consider **unigram character distributions** instead
  - Unigram means a single character; distribution means normalized counts
  - We count characters over all unique prefixes in the attack
  - We consider:
    - first character
    - all prefix characters

Comparing Three Attacks

Different Days -- Different Domains


Distribution of autoscout24.com on 8/24/18 representing cluster 0


Distribution of dollarshaveclub.com on 8/13/18

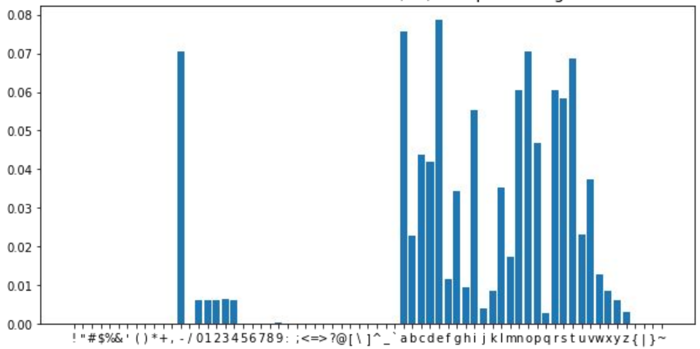

Character distribution of attack: 2018-08-13_shave.io cluster 0

Same day -- different domains

Distance for Character Distributions

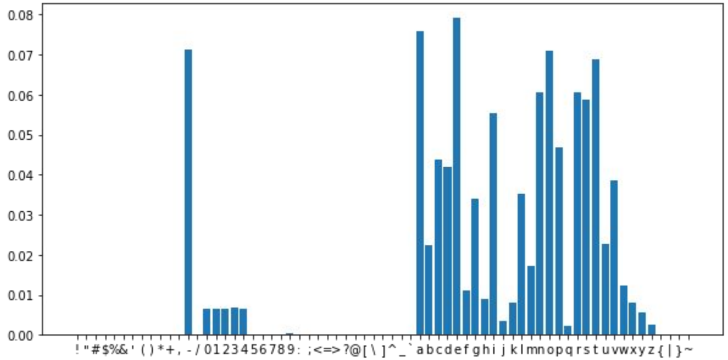
- Compute a distance measure between two distributions
  - There are several ways to use this
  - We use Jensen-Shannon Distance
- Calculate the distribution for every attack
- Calculate the pairwise distance for each attack

Distribution of autoscout24.com on 8/24/18 representing cluster 0



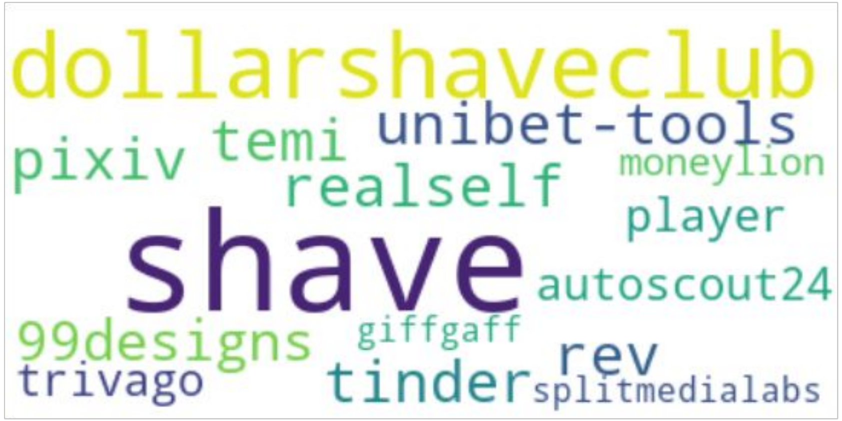
Small distance

Distribution of dollarshaveclub.com on 8/13/18

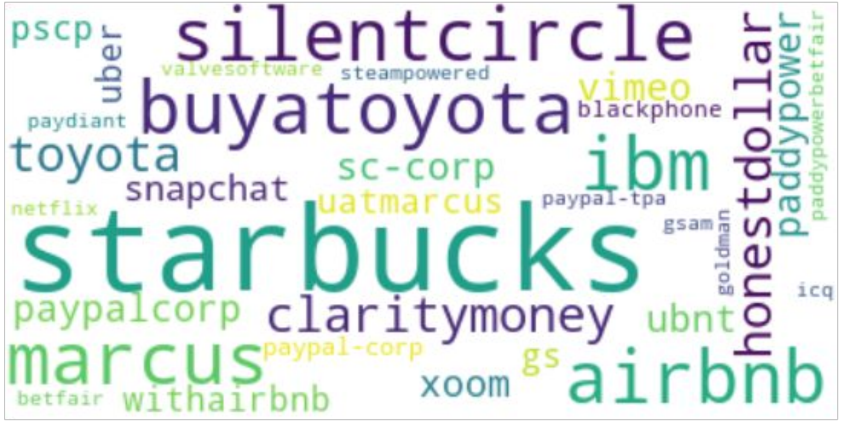


Attack Clusters From Unigram Distributions

domain Labels (second level) by Frequency in Cluster#0



domain Labels (second level) by Frequency in Cluster#2



These clusters cross many dates. Notice that we’ve picked up domains that are similar to each other solely by the single character distribution. This further supports that the attacks are related to each other.

What’s the goal?

We can’t compute pairwise distances for every attack over time, so what’s the point?

We hope to find a small number of **reference distributions** which we can use a measuring sticks for attacks.

- This means we really don’t care about clustering everything -- different than ‘typical’ clustering goals

Intuitively, this is creating a set of items to compare an unseen attack against, e.g., “oh, yeah, that attack is like airbnb on 7/31/18 attack”.

Reference Attacks

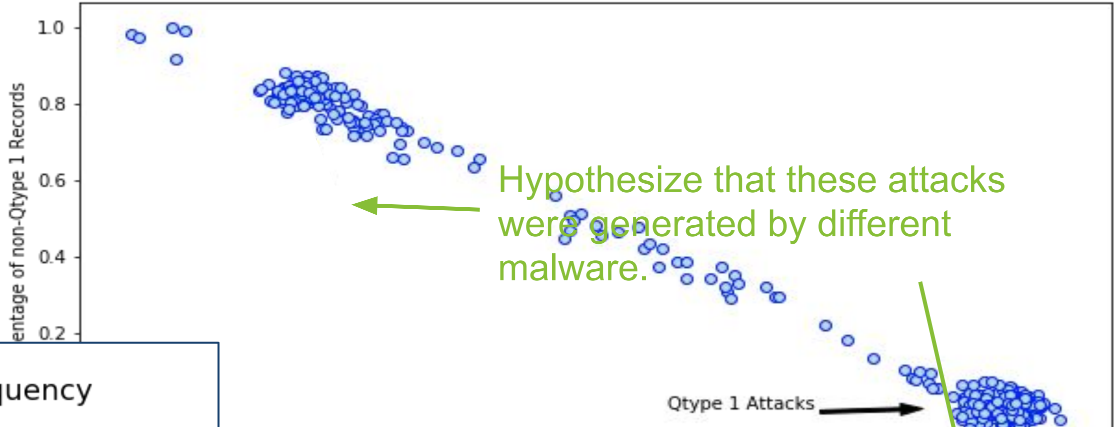
- Using data from June - August 2018, found 8 reference attacks
    - Used DBSCAN clustering and took the center attack
- We can now compare new attacks to these as one feature
- Found these distributions to still be consistent in 2019

Combine this character feature with others

Feature: Query Type

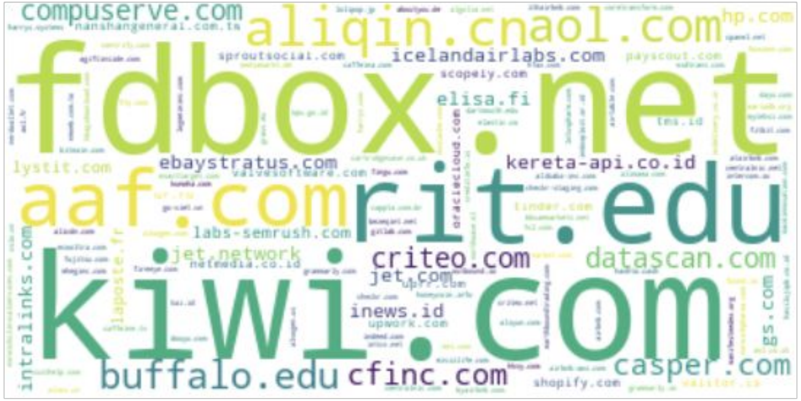
Comparison of Qtype 1 with Other Qtypes in Slow Drip Attacks

June - August 2018



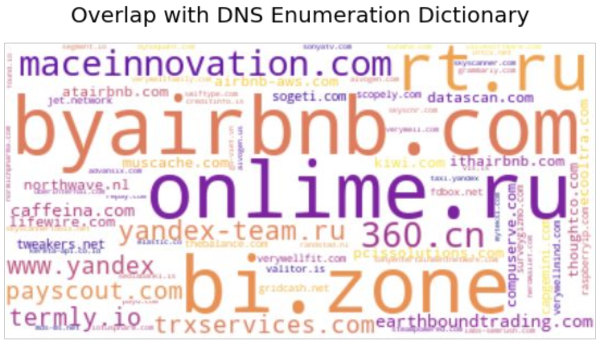
Hypothesize that these attacks were generated by different malware.

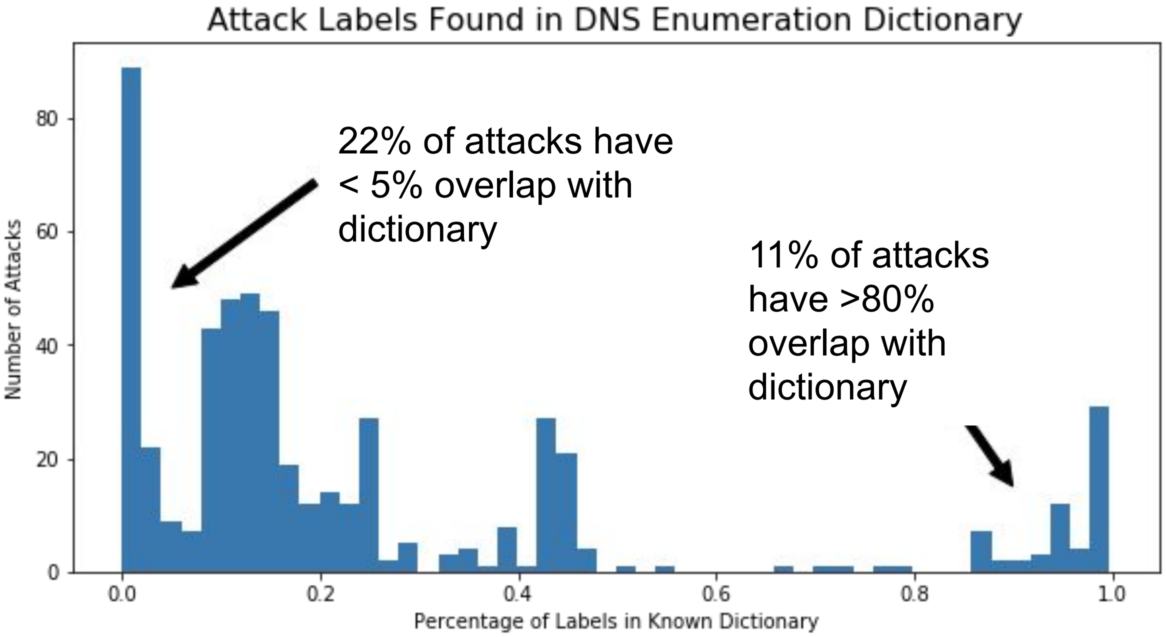
Attacks Using CNAME Query Types by Frequency



Feature: DNS Enumeration Dictionary

- A substantial number of the attacks use a dictionary for generating their random subdomain strings.


Overlap with DNS Enumeration Dictionary


Attack Labels Found in DNS Enumeration Dictionary

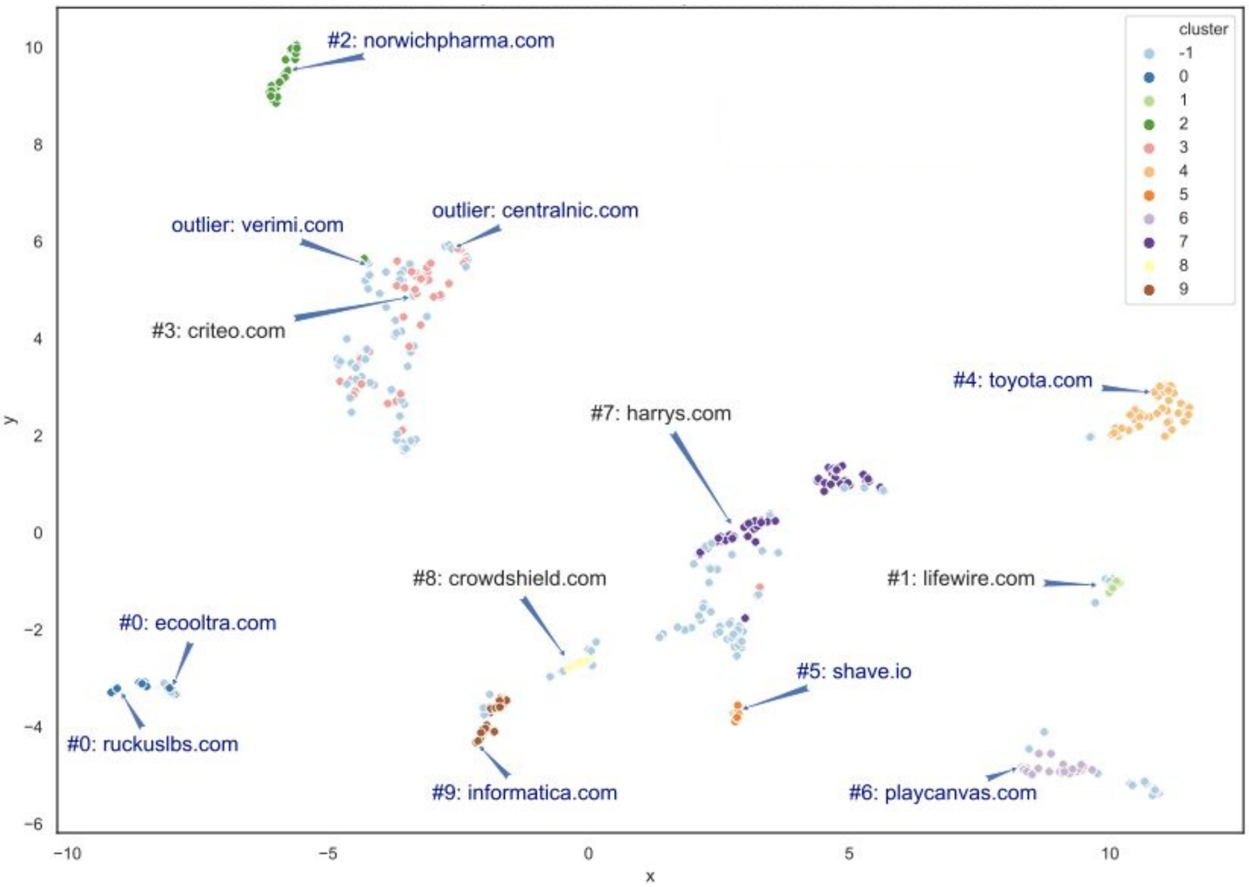
22% of attacks have < 5% overlap with dictionary

11% of attacks have >80% overlap with dictionary

Hypothesize that these attacks were generated by different malware.

Clustering the Attacks

- 20 Features in Total
  - Percentage of qtypes
  - Dictionary overlap
  - Character distributions
  - Time series
  - Percentage of unique labels in attack
  - Mean prefix lengths
  - etc.
- Used HDB-SCAN to cluster attacks in 20 dimensions
- Use UMAP to project 20-dimensional space into 2-d plane
  - -1 indicates an outlier



The Interesting Thing…

- Activity from 2014-2018 very high volume, one actor
  - Last seen May 2018
  - Mostly disrupted the middle Internet, not the targets
- Current activity is lower volume, harder to detect
  - Unclear what the motivations are
  - Totally different TTP than before
  - Not effective DDoS
  - Change in 2019
    - Weird mix of ‘target’ domains

Want more details?

Paper to be published in ACM Digital Threat: Research and Practice

Preprint available on Arxiv.org (outdated)

Data sample provided by Farsight Security.

- Contains a sample of queries from multiple attacks 2018 & 2019
- For personal research use; not to be disseminated or published without consent of Farsight Security
- `bit.ly/flocon20`

FARSIGHT SECURITY

Thank You
Vinaka
köszönöm
Kiitos
Maake
Asante
Shukria
Dhanyavadagalu
감사합니다
Dankscheen
спасибо
Kam Sah Hammida
شكرا
Manana
Dankon
Terima Kasih
Matondo
Dank Je
Mauruuru
Biyan
Diolch i Chi
Taiku
Tack
Blagodaram
Ngiyabonga
Dziekuje
Chokrane
Grazie
Juspaxar
Arigato
Mochchakkeram
Gracias
Paldies
Tingki
நன்றி
Bedankt
Dakujem
धन्यवाद
cảm ơn bạn
Kop Khun Khap
Ua Tsaug Rau Koj
Grazas
Gratias Tibi
Di Ou Mèsi
Kia Ora
Obrigado
Děkuji
Nirringrazzjak
ありがとう
Suksama
Matur Nuwun
Hvala
Rahmat
谢谢
хвала
Welalin
Djiere Dieuf
Misaotra
Danke
Merci
Go Raibh Maith Agat
Eskerrik Asko
Najis Tuke
Salamat
ขอบคุณ