Phishing URLs and Decision Trees

Hitesh Dharmdasani
Who am I?

• Cyber Crime, Internet threats, Malcode, Privacy, etc…

• GIT > George Mason > UC Berkeley > FireEye > With you

• Currently Informant Networks & Centre for Evidence Based Security Research

• http://hitesh.xyz
What is this talk about?

• Why solve this problem?

• What are the current measures to detect Phishing URLs

• What are decision trees?

• What is Spark? What is ML-Lib?

• Why use PySpark and MLLib?

• Some preliminary results

• Problems with this approach
What is this talk not about

- Success in Finding a solution
- Advice on Machine Learning
Phishing is an e-mail fraud scam conducted for the purposes of information or identity theft.

**Phishing** is the attempt to acquire sensitive information such as usernames, passwords, and credit card details (and sometimes, indirectly, money), often for malicious reasons, by masquerading as a trustworthy entity in an electronic communication.

[Phishing - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Phishing)
Why to solve this problem?

- The city of Belgaum gets 50 complaints of Phishing and Social Engineering every day. Think about bigger cities!
- A more real world problem in India than APT
- Takes advantage of the naive and gullible
Current Measures

• The ever prevalent blacklists

• Yara rules on E-mail bodies

• DMARC stops some spam. Not useful if email is from <INSERT COMPANY NAME>jobs2015@gmail.com. Hard to get right

• URL based features
  • Attributes are static

• Safe Browsing
  • Amazing effort! But only known pages classified. Also only with Chrome
  • Also, Not possible to integrate with private mail servers
Ground Zero


Decision Trees

- if-else statements generated by a computer
- Very powerful in expressing detection logic
- Human interpretation possible
  - Eliminate defects easily
- White box working
PySpark and ML-Lib

- PySpark is the Python Gateway to Apache Spark
- Allows for parallelism across any base data layer
- ML-Lib is a machine learning Library built in Spark. Leverages Scikit
- Allows to consume datasets that are distributed across a layer
- Apply's Algorithms over large datasets
Why use PySpark and MLLib

- Gathered dataset is 12 GB and growing per day, 2.6 lakh web pages
- Parsing HTML pages takes a while
- Don't want to roll out my own multiprocessing framework
- Can now export a model in PySpark. Yay!
- Ability to self learn. Data can grow
- I am biased towards Spark
What is the feature set

- Combining the best of both worlds
- Static Features - Depends on the URL
  - Is it a Dynamic DNS domain
  - Is it a direct IP address
  - Is this domain name using some branding
- Dynamic Features - Depends on HTML content
  - Does this page have a login form
    - Where does the form send its data
  - Is the form POSTing to some other domain?
MLLib and Decision Trees

- Features are represented as vector (RDD in Spark)
- `model.train` accepts the collection
- Classifier selects best feature to divide the training set at every iteration
- Iterate until feature set is best divided
- Not distinctive features are thrown away
What we got?

• A model with a training error of ~ 1%

• Higher False Positives with Benign Pages
  • Offset by whitelisting some domains, Ex: google.com
  • Problems due to lack of dataset

• Better results if we are cautious and incorporate existing whitelists and Safe Browsing
What works?

• Decision Trees quickly eliminate the obvious and nonsensical
  • Alexa does not help
  • Nor does DNS Reputation data
  • If some feature is equally likely in good and bad pages. Its useless
  • Since browsers get HTML anyways, Hook it up with an extension to call the API. Yay!
Problems

• Feature evasion
  • Attackers adding elements to HTML to evade features
• Needs active crawling. Not that big of a problem though
• Still not as good as a human looking at it. But things are getting better
• Spark does not have an API face
That’s all folks!

• Open to share the dataset and/or code

• Open sourced once it hits some measure of quality
  • Which you can help with ;)

• Happy to talk

Questions?

Email: hello@hitesh.xyz | Web: hitesh.xyz

Slides to talk on http://hitesh.xyz/files/phishing.pdf