WiMLDS Meetup – June 7, 2016

Deconstructing Domain Names to Reveal Latent Topics

Cheryl Flynn

AT&T Labs Research

Joint work with Kenny Shirley and Wei Wang

AT&T Labs Research, AT&T Security Research

Motivation

- Protecting web users from new malicious attacks is a critical task for network service providers
- Automatic methods for identifying new emerging malicious domains could help complement security monitoring efforts

Usefulness of higher-level characteristics

Known patterns among malicious domain names

- Phishing strategies – e.g., discount shopping, financial scams
- Typosquatting – e.g., www.att.com vs wwwatt.com
- Soundsquatting – e.g., bestbuy.com vs bestby.com

Research Questions

- Can we use topic models to identify interpretable latent structure in domain names?
- What does this structure reveal about the types of domain names visited within a network?
- Does this structure serve as a useful predictor for identifying emerging trends and threats?

Research Questions

- Can we use topic models to identify interpretable latent structure in domain names?
- What does this structure reveal about the types of domain names visited within a network?

Unsupervised Learning Approach

- Does this structure serve as a useful predictor for identifying emerging trends and threats?

Research Questions

- Can we use topic models to identify interpretable latent structure in domain names?
- What does this structure reveal about the types of domain names visited within a network?

Unsupervised Learning Approach

- Does this structure serve as a useful predictor for identifying emerging trends and threats?

Security Application

Review: Definition of a Domain Name

Full URL	http://www.more.example.com/path-to-url.html
Top level domain	http://www.more.example.com/path-to-url.html
Second-level domain	http://www.more.example.com/path-to-url.html

Domain name may only consist of:

- Alphanumeric characters
- Hyphens
- Top-level-domain (TLD)

Two Data Sources

1. Cellular data*

- Fresh domain data – domain names visited on a mobile device for the first time by any mobile user within 30 days
- Two weeks of data collected:
  - Thanksgiving week, 2013
  - Valentine's day week, 2014
- Roughly 800k unique domain names

2. DMOZ data

- Sample from Open Directory Project (DMOZ) in April 2015
- Roughly 1m unique domain names

*Data does not contain any additional information about domain traffic, nor does it contain any personal identifiers or search terms

Unsupervised learning approach – Step 1

Word Segmentation

- Dynamic programming algorithm (Norvig ‘09, Beautiful Data)

Sample DMOZ Domain Names

Domain Names	Segmented Text
flamingolake.com	flamingo, lake
tvguide.com	tv, guide
newadvent.org	new, advent
distinctivedirections.com	distinctive, directions
bellsofthesound.org	bells, of, the, sound

Unsupervised learning approach – Step 2

Topic Identification

- Biterm Topic Model (Yan et al., WWW2013)
  - Designed for short documents
  - Each biterm within a document has a latent topic assignment
  - Set of all biterms in the corpus is modeled as a mixture over topics

- BTM produced more coherent topics than competing methods including LDA and spherical K-means

Interpretable topics

Top 10 Most Coherent DMOZ Topics

Greek	Costa Rica	Hebrew	Christian	Animal Health
phi	costa	beth	holy	animal
sigma	rica	temple	lutheran	hospital
alpha	loss	shalom	trinity	clinic
gamma	weight	bnai	cross	vet
beta	contra	israel	blessed	veterinary

Los Angeles	Dog Breeds	Home Types	Golf	Fishing
angeles	short	log	golf	fishing
los	shepards	homes	club	fly
backers	hair	timber	course	fish
pike	german	cabin	disc	reels
speak	austrailian	cedar	tour	carp

Topics over time

Differences in topic distributions reflect holiday related topics and changes in topics due to what appears to be spam behavior

Most Significant Topic Differences

Topic	Z-score	Keywords	
2	11.61	friday, black, beats, 2013, dre	Holiday: Black Friday
9	-9.85	vow, chr, hear, here, reel	Possible spam behavior
42	7.94	tree, christmas, farm, trees, family	Holiday: Christmas Trees
19	7.09	outlet, kors, boots, sale, cheep	Possible spam behavior
1	6.83	monday, cyber, 2013, deals, ugg	Holiday: Cyber Monday
3	6.16	surv, lng, ys, you, nu	Possible spam behavior

Are the learned topics useful features for predicting malicious domain names?

Supervised learning application

Domain names labeled as malicious or benign based on rating reported by the website reputation site Web of Trust (www.mywot.com)

Data

- 167k unique domain names
- 15.7% of domain names labeled as malicious
- 80/20 training/test split

Detecting malicious domain names

Model

- Lasso regularized logistic regression
- Trained using 10-fold CV

Potential Sets Predictors

- Basic – presence of hyphens, digits, characters, TLD, etc.
- Words – individual words
- Topics – BTM learned topics, K = 50

Evaluation criteria

- Predictive accuracy
- Interpretability

Test-set performance

Same predictive accuracy with 1,300 fewer predictors

Model	AUC	# Potential Predictors	# Selected Predictors
Words + Basic	.797	30,819	5,551
Topics	.717	50	23
Words + Basic + Topics	.802	30,869	4,262

Most useful topic predictors

Most Malicious Topics

Topic	Keywords
26	sale, cheap, shoes, nike, outlet, 2014
27	sex, porn, tube, teen, girls, gay
33	payday, loan, credit, loans, cash, hour
35	top, online, credit, loans, cash, hour
12	my, free, 2, the, 4, web

Most Benign Topics

Topic	Keywords
7	county, of, city, chamber, society, hospital
50	creek, inn, lake, mountain, farm, river
25	club, north, coast, west, golf, lakes
37	and, photography, david, dr, photo, by
21	st, saint, parish, mary, marys, johns

- Highly interpretable
- Top 4 most malicious topics are related to discount shoes, adult content, financial scams, and drug/pharmaceutical offerings – all known phishing strategies
- Benign topics related to municipalities, geographical features, personal photography websites, and churches

Summary

Topic models can reveal meaningful and interpretable structure in domain names

Found that domain name topics can serve as useful features in predicting malicious domain names

Future work:

- Experiment with dynamic topic models
- Additional investigation into the usefulness of topics as features in supervised learning for malicious site detection