TEXAS A&M UNIVERSITY
Engineering

Click Trajectories: End-to-End Analysis of the Spam Value Chain

Presenter: Yu Shen

Outline

TEXAS A&M UNIVERSITY
Engineering

- Background
- Model
- Data Collection
- Analysis
- Review Discussion

Background

- Profitable industry – both spam and anti-spam
- Elements of the span value chain are studied in isolation

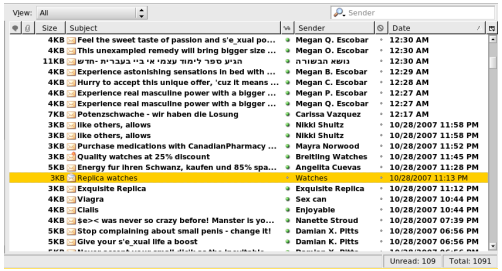
- A decade ago (about 2000) spammers might have handled virtually all aspects of the business including email distribution, site design, hosting, payment processing, fulfillment, and customer service.
- Today's spam business involves a range of players and service providers.

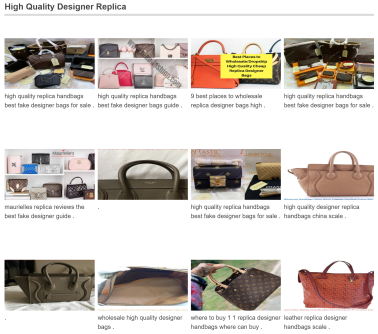
Modern Spam Working Model

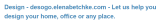
TEXAS A&M UNIVERSITY
Engineering

3 stages – pipeline

Advertising → Click Support → Realization







Modern Spam Working Model - Advertising

TEXAS A&M UNIVERSITY
Engineering

- All activities focused on reaching potential customers and enticing them into clicking on a particular URL

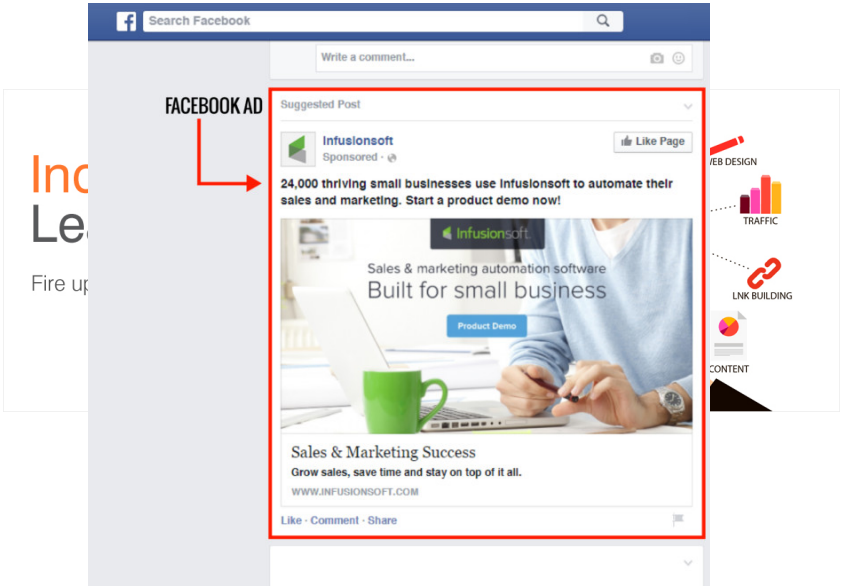
email spam

blog spam

Twitter spam

search engine optimization

sponsored advertising



Modern Spam Working Model - Advertising

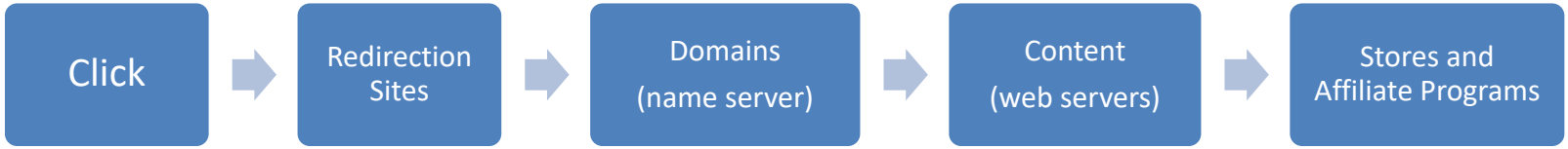
TEXAS A&M UNIVERSITY
Engineering

- Evolution of delivery and defense

Modern Spam Working Model – Click Support

TEXAS A&M UNIVERSITY
Engineering

- Direct user's browser to a Web site of interest, after clicking on the embedded URL



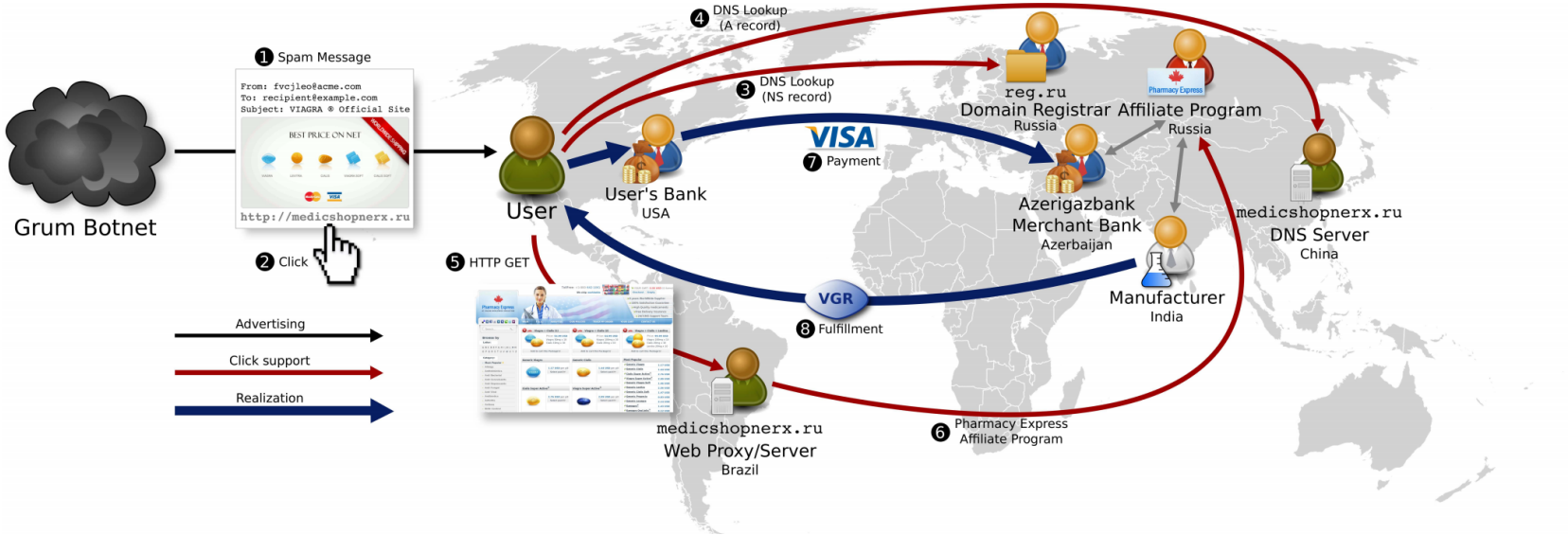
- Redirection sites (counter defensive measures)
  - Two ways: a legitimate third party | Manage DNS name resources by themselves

Modern Spam Working Model – Realization

- Payment services

- Fulfillment

Modern Spam Working Model – Example



Data Collection - Spam-Advertised URLs

TEXAS A&M UNIVERSITY
Engineering

- Duration: August 1, 2010 to October 31, 2010

Two sources:

Third party
including multiple commercial anti-spam providers

By botnet
from own botfarm environment

Feed Name	*Feed Description*	*Received URLs*	*Distinct Domains*
Feed A	MX honeypot	32,548,304	100,631
Feed B	Seeded honey accounts	73,614,895	35,506
Feed C	MX honeypot	451,603,575	1,315,292
Feed D	Seeded honey accounts	30,991,248	79,040
Feed X	MX honeypot	198,871,030	2,127,164
Feed Y	Human identified	10,733,231	1,051,211
Feed Z	MX honeypot	12,517,244	67,856
Cutwail	Bot	3,267,575	65
Grum	Bot	11,920,449	348
MegaD	Bot	1,221,253	4
Rustock	Bot	141,621,731	13,612,815
Other bots	Bot	7,768	4
Total		968,918,303	17,813,952

Data Collection - Crawler data

TEXAS A&M UNIVERSITY
Engineering

Two parts: DNS & Web

DNS Crawler:

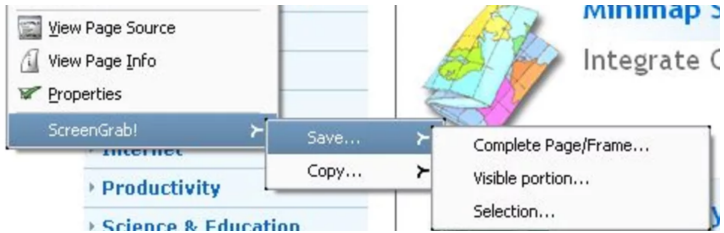
- Difficulty: fast flux - minimize central points of weakness
  - Solution: *X. Hu, M. Knysz, and K. G. Shin. RB-Seeker: Auto-detection of Redirection Botnets. In Proc. of 16th NDSS, 2009.*
- query servers repeatedly to enumerate the set of domains collectively used for click support

Data Collection - Crawler data

TEXAS A&M UNIVERSITY
Engineering

Web Crawler:

- replicates the experience of a user clicking on the URLs derived from the spam feeds
- captures any application-level redirects (HTML, JavaScript, Flash)
- Screenshots by Screengrab! Extension



- Firefox working parallelly fetching the web contents

Data Collection - Crawler data

TEXAS A&M UNIVERSITY
Engineering

Result: nearly complete coverage: crawls over 98% of the URLs received

Stage	*Count*	
Received URLs	968,918,303	
Distinct URLs	93,185,779	(9.6%)
Distinct domains	17,813,952	
Distinct domains crawled	3,495,627	
URLs covered	950,716,776	(98.1%)

Notes: crawl URLs that account for only 20% of the whole

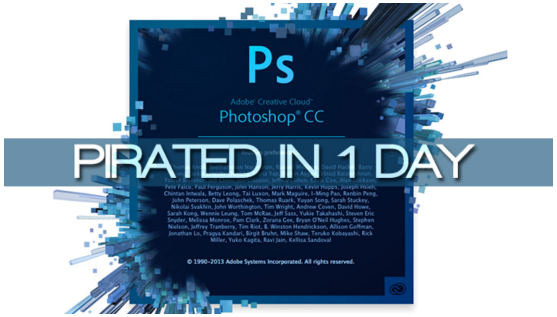
Explanation: 80% of domains, and the corresponding 2% URLs are from the domain-poisoning spam (Rustock) - no not reflect real sites

Data Collection - Content Clustering & Tagging

TEXAS A&M UNIVERSITY
Engineering

Focus on 3 types of products:

`pharmaceuticals` `replicas` `software`



Methodology: match sites with lexically similar content structure

Data Collection - Content Clustering & Tagging

TEXAS A&M UNIVERSITY
Engineering

Step 1: Content clustering

- Basis: HTML text
- Approach: q-gram similarity
- Fingerprint: a set of multiple independent hash values over all 4-byte tokens of the HTML text

Q-gram language models:

This is Big Data AI Book

Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

Data Collection - Content Clustering & Tagging

TEXAS A&M UNIVERSITY
Engineering

Step 2: Category tagging

- Separates clusters that sells goods we want
- Avoid false negative: be intentionally conservative
  - potentially including clusters that turn out to be false positives to ensure that we include all clusters that fall into one of our categories
- identify interesting clusters by generic keywords
- Manual work: check for false negatives
  - Randomly select 1,000 pages from 675 untagged clusters, but find none missed

Data Collection - Content Clustering & Tagging

TEXAS A&M UNIVERSITY
Engineering

Step 3: Program tagging

- Label clusters with specific program tags
  - E.g. distinct storefront brands of EvaPharmacy
- Methodology – vague…

Data Collection - Content Clustering & Tagging

TEXAS A&M UNIVERSITY
Engineering

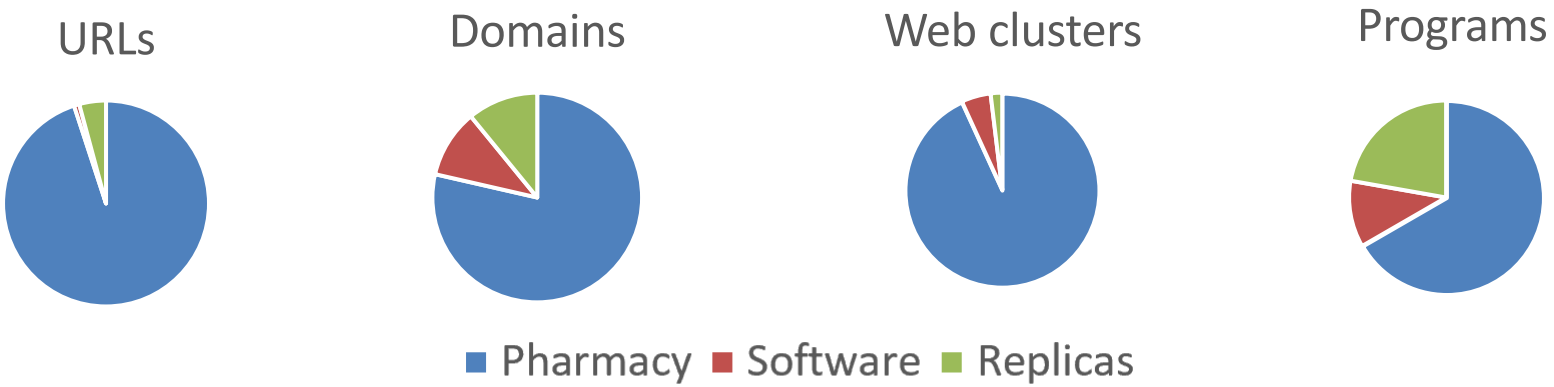
Affiliate Program		*Distinct Domains*	*Received URLs*	*Feed Volume*
RxPrm	RX–Promotion	10,585	160,521,810	24.92%
Mailn	Mailien	14,444	69,961,207	23.49%
PhEx	Pharmacy Express	14,381	69,959,629	23.48%
EDEx	ED Express	63	1,578	0.01%
ZCashPh	ZedCash (Pharma)	6,976	42,282,943	14.54%
DrMax	Dr. Maxman	5,641	32,184,860	10.95%
Grow	Viagrow	382	5,210,668	1.68%
USHC	US HealthCare	167	3,196,538	1.31%
MaxGm	MaxGentleman	672	1,144,703	0.41%
VgREX	VigREX	39	426,873	0.14%
Stud	Stud Extreme	42	68,907	0.03%
ManXt	ManXtenz	33	50,394	0.02%
GlvMd	GlavMed	2,933	28,313,136	10.32%
OLPh	Online Pharmacy	2,894	17,226,271	5.16%
Eva	EvaPharmacy	11,281	12,795,646	8.7%
WldPh	World Pharmacy	691	10,412,850	3.55%
PHOL	PH Online	101	2,971,368	0.96%
Aptke	Swiss Apotheke	117	1,586,456	0.55%
HrbGr	HerbalGrowth	17	265,131	0.09%
RxPnr	RX Partners	449	229,257	0.21%
Stmul	Stimul-cash	50	157,537	0.07%
Maxx	MAXX Extend	23	104,201	0.04%
DrgRev	DrugRevenue	122	51,637	0.04%
UltPh	Ultimate Pharmacy	12	44,126	0.02%
Green	Greenline	1,766	25,021	0.36%
Vrlty	Virility	9	23,528	0.01%
RxRev	RX Rev Share	299	9,696	0.04%
Medi	MediTrust	24	6,156	0.01%
ClFr	Club-first	1,270	3,310	0.07%
CanPh	Canadian Pharmacy	133	1,392	0.03%
RxCsh	RXCash	22	287	<0.01%
Staln	Stallion	2	80	<0.01%
	Total	54,220	346,993,046	93.18%
Royal	Royal Software	572	2,291,571	0.79%
EuSft	EuroSoft	1,161	694,810	0.48%
ASR	Auth. Soft. Resellers	4,117	65,918	0.61%
OEM	OEM Soft Store	1,367	19,436	0.24%
SftSl	Soft Sales	35	93	<0.01%
	Total	7,252	3,071,828	2.12%
ZCashR	ZedCash (Replica)	6,984	13,243,513	4.56%
UltRp	Ultimate Replica	5,017	10,451,198	3.55%
Dstn	Distinction Replica	127	1,249,886	0.37%
Exqst	Exquisite Replicas	128	620,642	0.22%
DmdRp	Diamond Replicas	1,307	506,486	0.27%
Prge	Prestige Replicas	101	382,964	0.1%
OneRp	One Replica	77	20,313	0.02%
Luxry	Luxury Replica	25	8,279	0.01%
AffAc	Aff. Accessories	187	3,669	0.02%
SwsRp	Swiss Rep. & Co.	15	76	<0.01%
WchSh	WatchShop	546	2,086,891	0.17%
	Total	7,530	15,330,404	4.73%
	Grand Total	69,002	365,395,278	100%

Data Collection - Content Clustering & Tagging

TEXAS A&M UNIVERSITY
Engineering

- Overall result

Stage	*Pharmacy*	*Software*	*Replicas*	*Total*
URLs	346,993,046	3,071,828	15,330,404	365,395,278
Domains	54,220	7,252	7,530	69,002
Web clusters	968	51	20	1,039
Programs	30	5	10	45

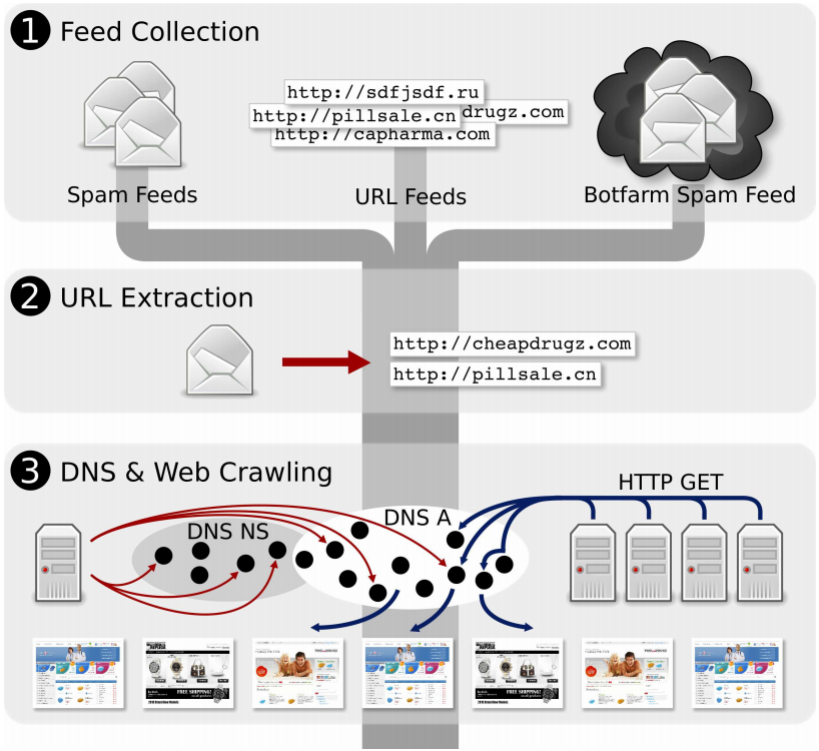


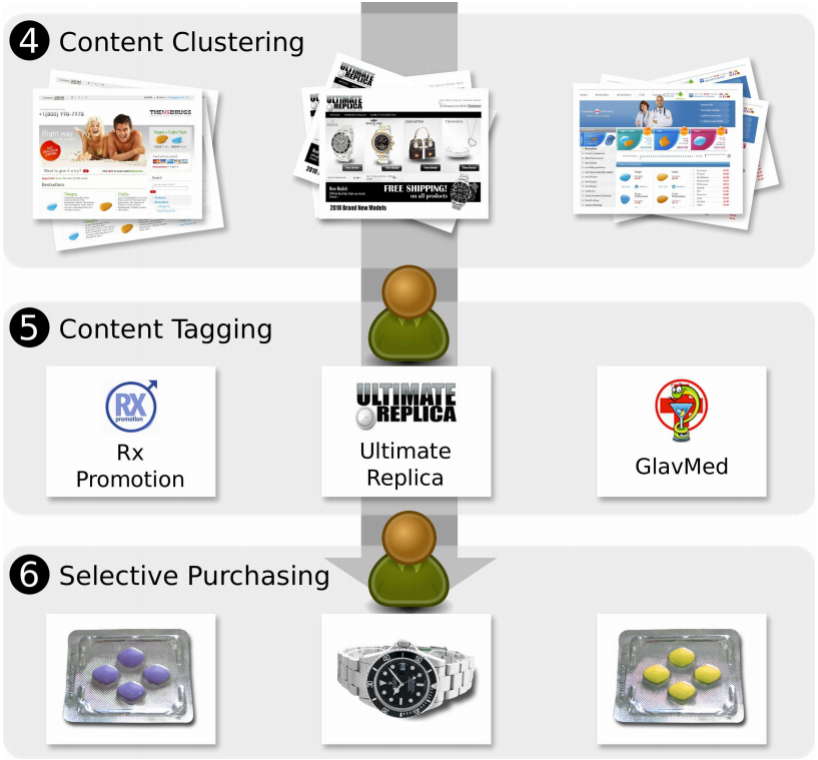
Data Collection - Purchasing

TEXAS A&M UNIVERSITY
Engineering

- Goal: identify differences or similarities in suppliers
- Basis: contents (e.g., lot numbers) and packaging (nominal sender, packaging type, etc.)
- Method: place multiple purchases from each major affiliate program or store “brand”, ordering the same “types” of product from different sites
- Process: attempt 120 purchases, 76 authorized, 56 settled

Data Collection - Overview



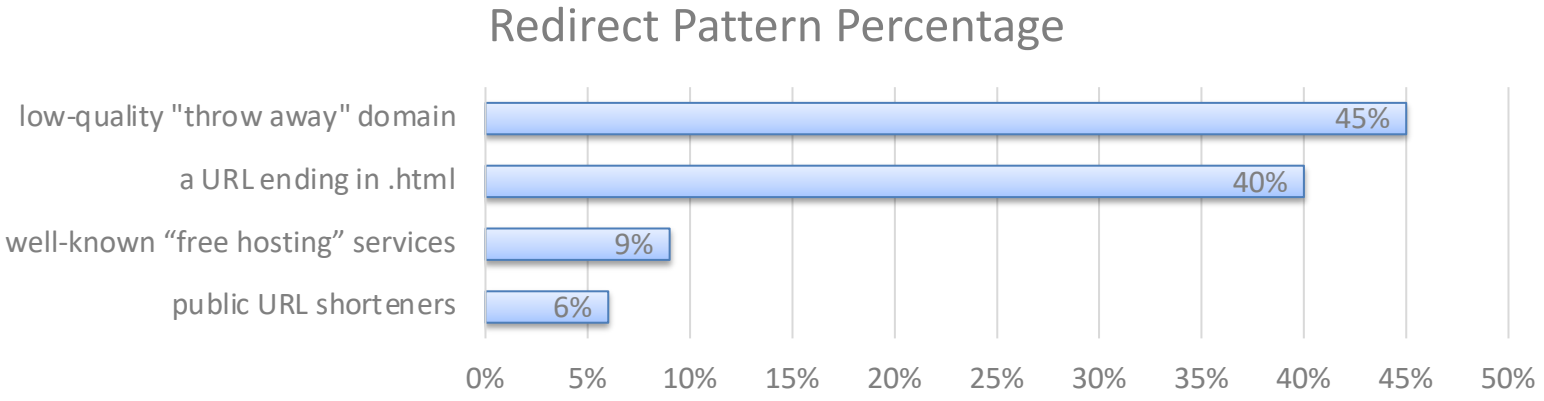


Analysis

TEXAS A&M UNIVERSITY
Engineering

- Major goal: identify any “bottlenecks” in the spam value chain

- [problem] Redirection in click support: use the final domain
  - URLs using redirection: 32%

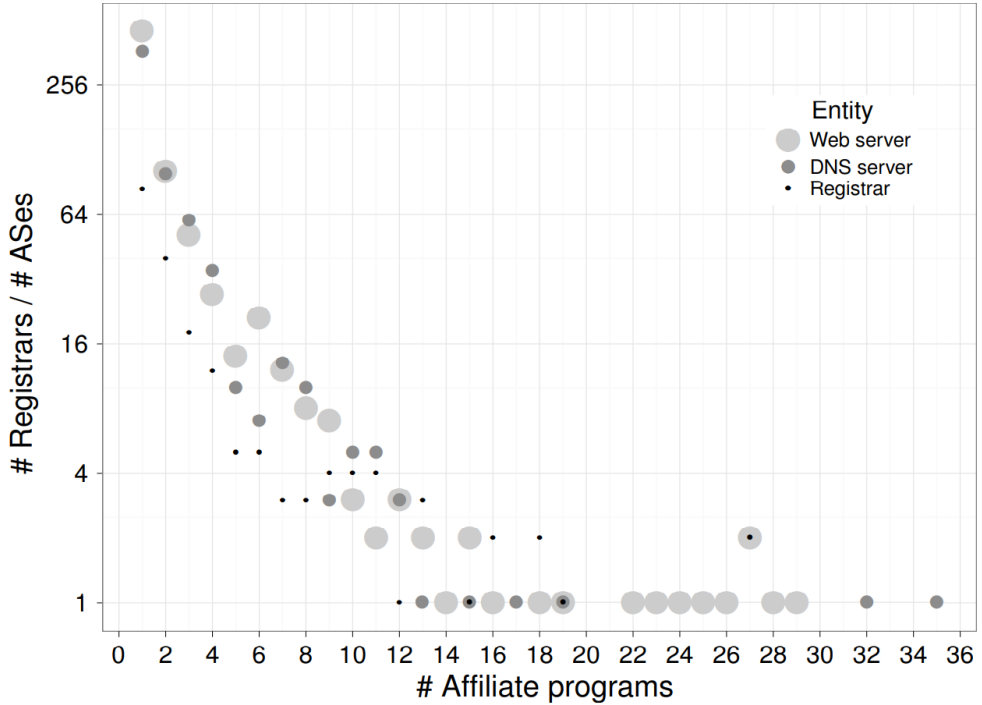


Analysis

TEXAS A&M UNIVERSITY
Engineering

- examine the set of resources used by sites for each affiliate program
- Network infrastructure sharing (at least 1):
  - Domain name
  - Registrar
  - Provision server

Analysis - Click Support

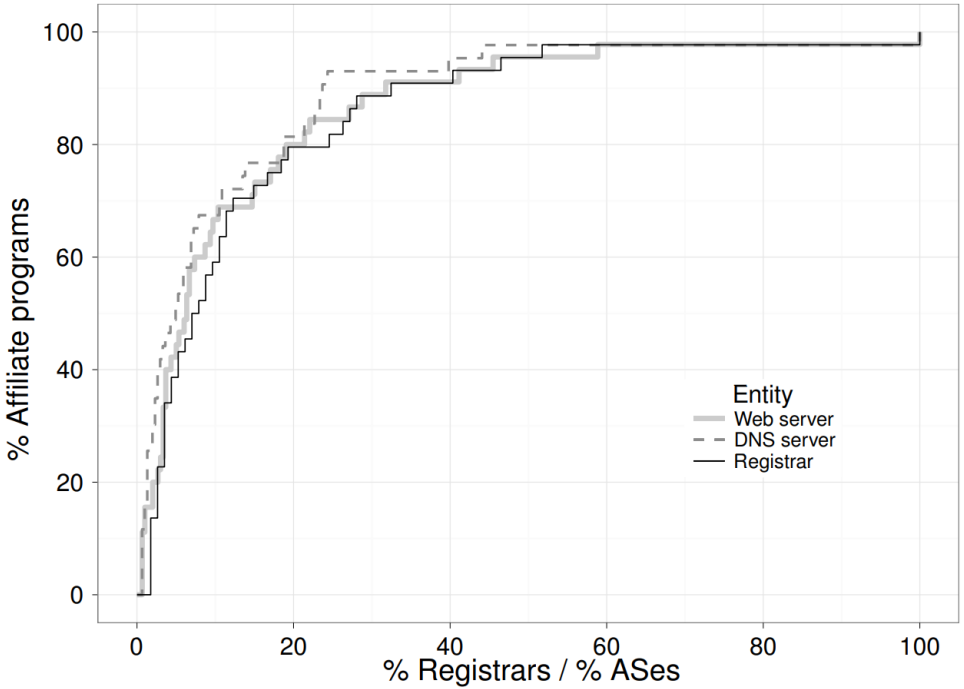


Network infrastructure sharing among affiliate programs is concentrated in a small number of registrars and Autonomous Systems (ASes)

Analysis - Click Support

TEXAS A&M UNIVERSITY
Engineering

Distribution of infrastructure among affiliate programs



Analysis - Realization

TEXAS A&M UNIVERSITY
Engineering

Hypothesis

- Realization infrastructure is the province of affiliate programs

Fact

- Consistency in payment processing and fulfillment between different instances

Support

- Email templates & Order number formats

Analysis - Realization

TEXAS A&M UNIVERSITY
Engineering

Merchant banks

Bank Name	*BIN*	*Country*	*Affiliate Programs*
Azerigazbank	404610	Azerbaijan	GlvMd, RxPrm, PhEx, Stmul, RxPnr, WldPh
B&N	425175	Russia	ASR
B&S Card Service	490763	Germany	MaxGm
Borgun Hf	423262	Iceland	Trust
Canadian Imperial Bank of Commerce	452551	Canada	WldPh
Cartu Bank	478765	Georgia	DrgRev
DnB Nord (Pirma)	492175	Latvia	Eva, OLPh, USHC
Latvia Savings	490849	Latvia	EuSft, OEM, WchSh, Royal, SftSl
Latvijas Pasta Banka	489431	Latvia	SftSl
St. Kitts & Nevis Anguilla National Bank	427852	St. Kitts & Nevis	DmdRp, VgREX, Dstn, Luxry, SwsRp, OneRp
State Bank of Mauritius	474140	Mauritius	DrgRev
Visa Iceland	450744	Iceland	Staln
Wells Fargo	449215	USA	Green
Wirecard AG	424500	Germany	ClFr

Analysis - Realization

TEXAS A&M UNIVERSITY
Engineering

Product Suppliers

Supplier	*Item*	*Origin*	*Affiliate Programs*
Aracoma Drug	Orange bottle of tablets (pharma)	WV, USA	ClFr
Combitic Global Caplet Pvt. Ltd.	Blister-packed tablets (pharma)	Delhi, India	GlvMd
M.K. Choudhary	Blister-packed tablets (pharma)	Thane, India	OLPh
PPW	Blister-packed tablets (pharma)	Chennai, India	PhEx, Stmul, Trust, ClFr
K. Sekar	Blister-packed tablets (pharma)	Villupuram, India	WldPh
Rhine Inc.	Blister-packed tablets (pharma)	Thane, India	RxPrm, DrgRev
Supreme Suppliers	Blister-packed tablets (pharma)	Mumbai, India	Eva
Chen Hua	Small white plastic bottles (herbal)	Jiangmen, China	Stud
Etech Media Ltd	Novelty-sized supplement (herbal)	Christchurch, NZ	Staln
Herbal Health Fulfillment Warehouse	White plastic bottle (herbal)	MA, USA	Eva
MK Sales	White plastic bottle (herbal)	WA, USA	GlvMd
Riverton, Utah shipper	White plastic bottle (herbal)	UT, USA	DrMax, Grow
Guo Zhonglei	Foam-wrapped replica watch	Baoding, China	Dstn, UltRp

Analysis - Intervention analysis

From the point of defender: which intervention has the most impact?

- blocking its advertising
  - filtering spam
- disrupting its click support
  - takedowns for name servers of hosting sites
- interfering with the realization step
  - shutting down merchant accounts

Analysis - Intervention analysis

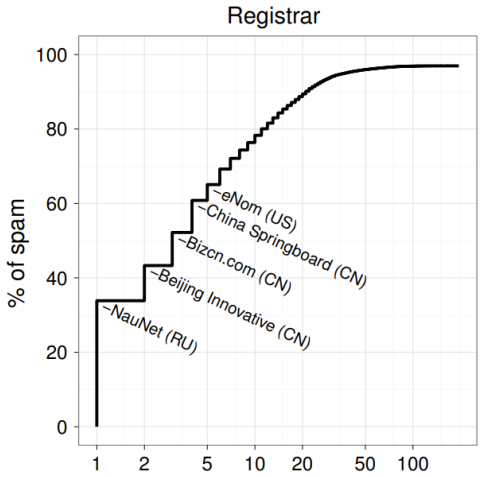
TEXAS A&M UNIVERSITY
Engineering

- Strategy Evaluation
  1. overhead to implement
  2. Business impact (replacement cost & opportunity cost)

❖ Registrar

significant concentrations among the top few providers

takedowns would seem to be an effective strategy



Analysis - Intervention analysis

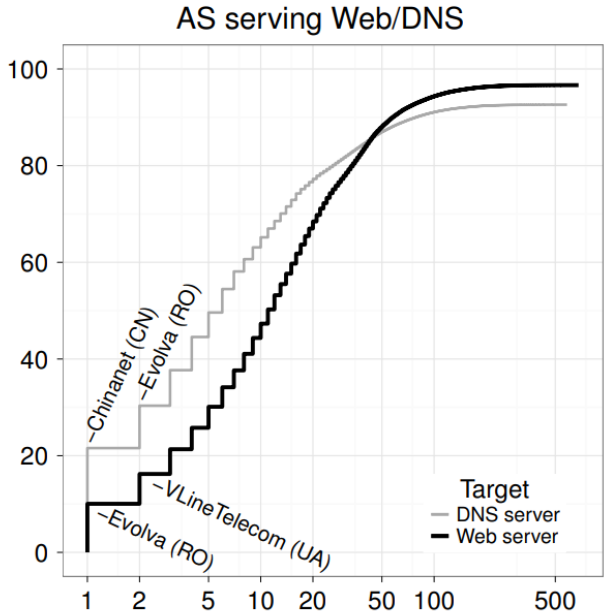
TEXAS A&M UNIVERSITY
Engineering

- Hosting

Analysis target: the number of distinct ASs that would need to be contacted

“worst case” model - that all such resources must be eliminated

Similar concentrations like registrar



Analysis - Intervention analysis

TEXAS A&M UNIVERSITY
Engineering

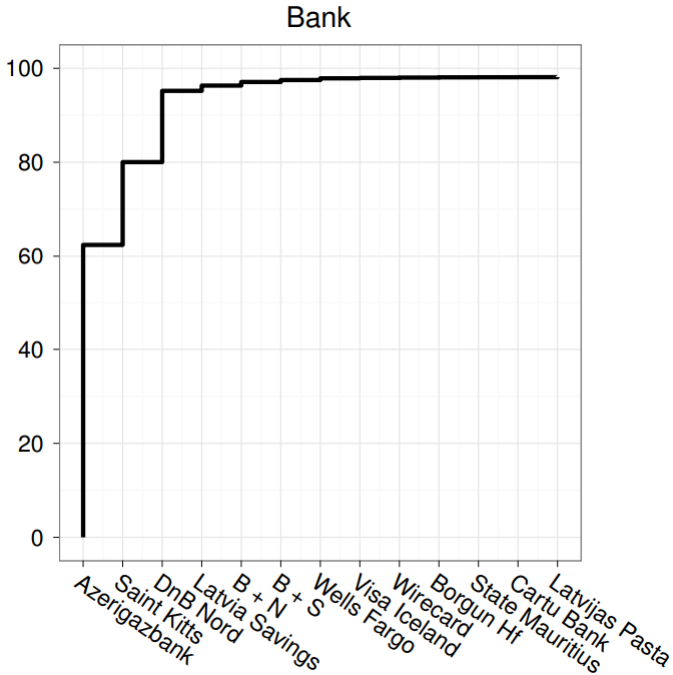
- Payments

Not buy from all domains

Assumption: bank use will be consistent across domains belonging to the same brand or affiliate program

95% transactions are with 3 banks:

- Azerigazbank
- Saint Kitts
- Latvia Savings



Analysis - Intervention analysis

TEXAS A&M UNIVERSITY
Engineering

- availability of alternatives & switching cost

	Alternatives	Switching Cost
Hosting	Vast (>1M)	Low (almost none)
Registrar	Few (roughly 900 gLTD registrars)	Low (spammers show great agility)
Bank	Few (a small part of the banks accept processing high-risk transactions)	High (setup fees, legitimate merchant account)

Analysis – Policy Options

TEXAS A&M UNIVERSITY
Engineering

- Directly pressure the merchant banks to stop doing business with such merchants
  - Slow
  - Incongruities in intellectual property protection

- refuse to settle certain transactions with the spam-supporting banks
  - Update quickly
  - Legal basis

Review

TEXAS A&M UNIVERSITY
Engineering

❖ Merit - Good overview analysis

Reviewer1

- The paper provides a big picture of the spam ecosystem, which will clearly benefit future research in this direction.
- The authors divide the spam chain to advertising, click support and realization, providing a comprehensive understanding of the spam business.

Reviewer2

- Comprehensive thinkings. First full analysis on the entire value chain.

Reviewer3

- The paper brought me a full view of the spam ecosystem. Even if I'm not familiar with how spam works, I can still understand the paper to some degree.

Review

TEXAS A&M UNIVERSITY
Engineering

❖ Merit – Rich background info & graphics

Reviewer4

- This paper was very well written and thought out. One of the primary strengths is the detailed background information. … Figure 2 really was set up very well to display step by step how the data collection and processing workflow occurred.

Reviewer5

- Specifically, the beginning sections discussing the background research and related work goes in depth about the step by step implementation of spam systems and how they try to target their customers.

Reviewer6

- The graphic well shows how much of a global industry this has become hardening attempts for one government to stop this industry.

Review

TEXAS A&M UNIVERSITY
Engineering

❖ Weakness – Lack of detailed explanation on techniques

Reviewer3

- For some techniques mentioned in this paper like maximum likelihood methods and q-gram, I would like to see the more detailed examination of how they using it to process data.

Reviewer7

- The approach for content clustering is out-of-date. We have many more powerful tools today.
- In conclusion, they mentioned that they "described a framework for ...". But the framework is not clear in the paper.

Reviewer8

- When classifying clusters, if using image recognition to categorize three kinds of web sites, personally assumed, would be better than HTML contents analysis. At least untagged groups would be fewer.

Review

TEXAS A&M UNIVERSITY
Engineering

❖ Different Views: Is payment the bottleneck? / Is the policy practical?

Reviewer1

- Some of the proposed intervening measures are hard to realize. For example, one approach for intervening at the payment tier of the value chain suggested in the paper is to directly engage the merchant banks and pressure them to stop doing business with such merchants. This measure needs support from the government. It is much more difficult to realize than designing the anti-spam system.

Reviewer8

- The conclusion, "payment" has few connections with prior research approaches, classifying, content analysis. This makes the whole paper a bit waste of efforts.

Reviewer6

- With the use of a VISA card issuer that is willing to help provide non traditional amount of information to the study I would have liked for the authors to ask the issuer directly how they feel about any of the policy recommendations.

Review

TEXAS A&M UNIVERSITY
Engineering

❖ Other questions:

Reviewer1

- The data collection extracts URL. However, some spammers may not use the URL. They may only use contact information like a phone number.

Reviewer7

- The paper is not well-organized. From their analysis, we can get gradually understand the spam value chain. They illustrated each stage very clearly. But they missed an overview of the chain. It would be better if they can spend some words on the overview of the chain.

TEXAS A&M UNIVERSITY

Engineering