

Methodological issues in:

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

Stefan Savage
University of California, San Diego

Click Trajectories paper

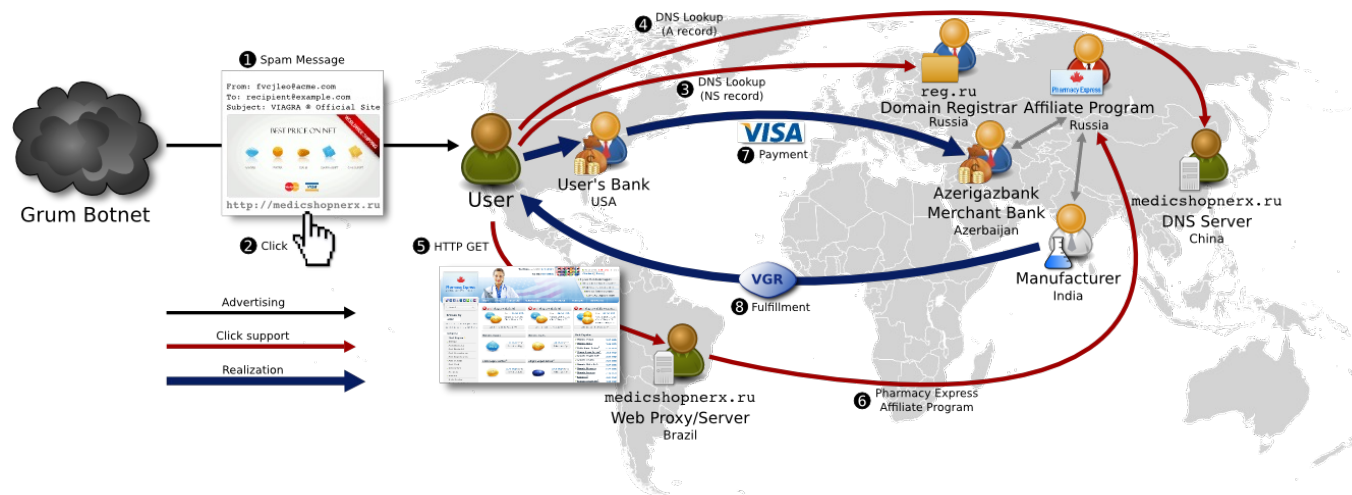
Motivation

- What is underlying motivation for this paper? What do they want to find out?

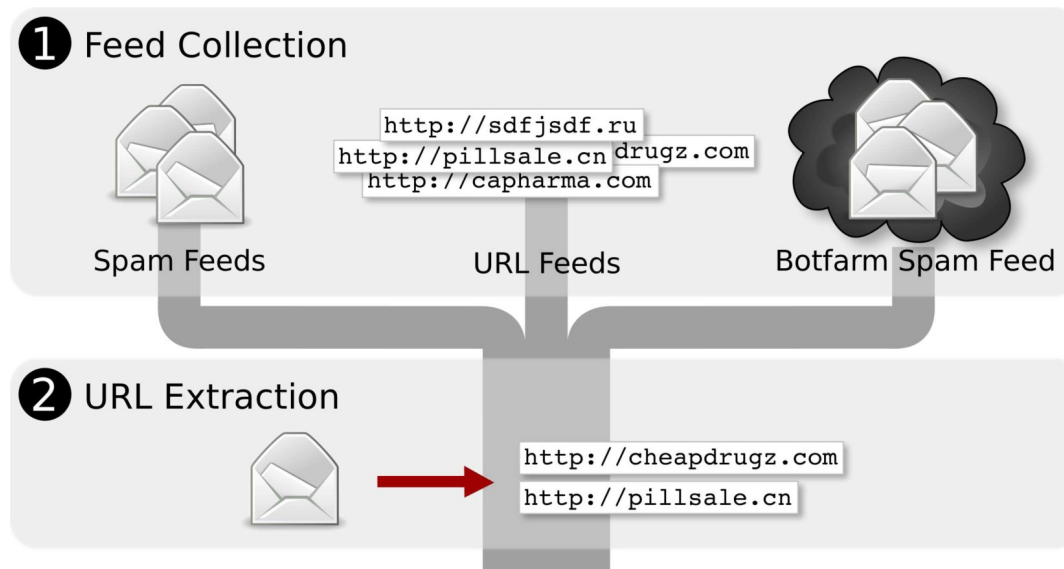
Note grounding in economic analysis

- “Spam... is a business”
- Goal is to understand something about business model
 - What is “supply chain” for spammers (i.e., what do they need?)
 - Is there aggregation/consolidation in parts of this supply chain (i.e., such that undermining a resource would have disproportionate impact)

Lifecycle of a spam e-mail "trajectory"



Measurement stages



Let's start with the spam

12 different feeds

- Collected different ways
 - MX honeypot
 - Seeded honey accounts
 - Human identified
 - Bot output
- Bias issues?
 - Why 141M URLs and 13M domains from Rustock?
 - Why is URL/domain ratio so variable?

Is this representative?

- What do I actually care about?
- Unique URLs, Unique domains? Something else?
- Do we have the same kinds of problems here we had in the TLS paper?

Was this level of data collection necessary?

- Maybe not, but wasn't clear at the time
- See "Taster's Choice: A Comparative Analysis of Spam Feeds", IMC '12

<i>Feed Name</i>	<i>Feed Description</i>	<i>Received URLs</i>	<i>Distinct Domains</i>	URLs/domains
Feed A	MX honeypot	32,548,304	100,631	323
Feed B	Seeded honey accounts	73,614,895	35,506	2073
Feed C	MX honeypot	451,603,575	1,315,292	343
Feed D	Seeded honey accounts	30,991,248	79,040	392
Feed X	MX honeypot	198,871,030	2,127,164	93
Feed Y	Human identified	10,733,231	1,051,211	10
Feed Z	MX honeypot	12,517,244	67,856	184
Cutwail	Bot	3,267,575	65	50270
Grum	Bot	11,920,449	348	34254
MegaD	Bot	1,221,253	4	305313
Rustock	Bot	141,621,731	13,612,815	10
Other bots	Bot	7,768	4	1942
Total		968,918,303	17,813,952	

Table I: Feeds of spam-advertised URLs used in this study. We collected feed data from August 1, 2010 through October 31, 2010.

heavily skewed in their makeup. For example, we received over 11M URLs from the Grum bot, but these only contained 348 distinct registered domains. Conversely, the 13M distinct domains produced by the Rustock bot are artifacts of a "blacklist-poisoning" campaign undertaken by the bot operators that comprised millions of "garbage" domains [54]. Thus, one must be mindful of these issues when analyzing such feed data in aggregate.

Why just focus on the URLs?

Why don't we care about the content of the message?

- Text, images, etc

Do we want all URLs in e-mails or only some of them?

What are opportunities for error here?

- False positives (looking at sites that aren't being advertised by spam)
- False negatives (missing sites that are spam-advertised)

Getting data via crawling

From a list of URLs

- Extract the domain name part and look it up in DNS
 - i.e., what resources are needed to “resolve” it to the site’s IP address
- Visit the Web site itself
 - Grab the Web page content (DOM) and a screen shot

Why isn’t this easy?

- Main reason: adversarial behavior
- Lots of anti-spam efforts
 - Blocking email, taking down sites, blocking web visits, etc
 - So there is spammer interest in making it complex
- Secondary reason: crawling has lots of details to get right

DNS crawler

Simple version:

- Take URL http://bar.foo.com:/stuff/?q=viagra&stuff_and=a6857920e519odbo&ei=dry677
- Extract domain (bar.foo.com); Go to DNS and lookup A record (i.e., IP address)
- Why isn't this enough?

First, we want all the **pieces of infrastructure**

- NS record (what name server is used)
- SOA record (start of authority; who is responsible for this domain?)
- CNAME record (if it forwards lookups somewhere else)
- Other records in case they tell us something (MX – mail, TXT – misc)

Second, **answers may be inconsistent** to “hide” from investigators

- Fast Flux: technique whereby mapping from NS -> A record changes constantly
- Double Flux: technique whereby mapping from domain -> NS changes as does NS -> A
- How does DNS crawler manage this risk?

Aside: not sure all this was necessary given ultimate direction of paper

Web crawler

Parallel crawler

- ~40 servers, each running 100 instances of Firefox

Why not use wget or curl? Or selenium?

- Forwarding – so many kinds (32% redirect!)
 - Shorteners, freehosting, compromised doorway pages
 - Many mechanisms
 - HTTP redirect (30x), Javascript, Flash
- Javascript impact
 - Can change DOM; multi-stage download
- Screenshot (need to render page!)

Wouldn't they notice millions of queries coming from UCSD?

- Yes, so crawlers "tunnel" traffic via other non-UCSD IPs
 - "Stargate" service developed for this project

⁴Among the complexities, scammers are aware that security companies crawl them and blacklist IP addresses they suspect are crawlers. We mitigate this effect by tunneling requests through proxies running in multiple disparate IP address ranges.

Web crawling

Only crawl subset of URLs

- Prioritize new registered domains. Why?
- Rate limit URLs to same registered domain. Why?

Why not simply crawl every URL?

- Load – at 30sec per page, that would be over 8M hours of crawling
- Limited value (most URLs are replicas)
- Mitigate detection risk (don't want crawler blocklisted)
- Want to grab new sites before they disappear!

Significant reduction

- 968 URLs -> visit only ~15M URLs
- Is this ok?

<i>Stage</i>	<i>Count</i>	
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%)

Table II: Summary results of URL crawling. We crawl the registered domains used by over 98% of the URLs received.

Aside: affiliate program structure (common to much of cybercrime)

Division of labor

- **Affiliates** handle customer acquisition (e.g., spam, SEO)
 - Independent contractors
 - Paid 25-50% commission for successful sale
- **Affiliate programs** handle business processes
 - Payment processing, customer service, fulfillment
 - Sometimes content hosting and domain name registration

Why?

- Risk transfer (for both parties)
- Specialization lowers cost structure

[Наши сайты](#)[Преимущества](#)[Вопрос - ответ](#)[Что тут происходит](#)[Контакты](#)[Напомнить пароль](#)

Добро пожаловать на RX-Promotion

RX-Promotion — партнерская программа, позволяющая участникам зарабатывать на продаже медикаментов там, где они хотят и так, как они хотят. Можно открыть собственную онлайн-аптеку или работать через нашу сеть.

Главное — это найти покупателя и продать медикаменты. Ну и, конечно, не забыть получить прибыль: вплоть **до 60% от сделки**. Считаете это предложение невыгодным? Что ж, остается вам только позавидовать.

Кстати, а сейчас мы разыгрываем [Золотой Слиток](#).



Тех.отдел

Общие вопросы

ICQ: 402961146

ICQ: 457098148

Тех. вопросы

ICQ: 50423090

Еще ICQ



Skype:

rx-promotion

[РЕГИСТРАЦИЯ ►](#)

У нас отлично:



— Конкурсы? Ну и что?
— Действительно, ну и что,
подумаешь, Харлей выиграть.

**КОМИССИЯ
ДО 60%
ПО РЕФЕРАЛАМ
ДО 15%**



RCE

Major pharmaceutical affiliate programs in 2011



Organizing the sites

Content clustering applied to Web content

- Lots of different URLs/domains point to similar/identical sites

Why?

- Cluster on Q-gram similarity of DOM (i.e., HTML) content – 4-byte strings
Jaccard similarity index (basically intersection divided by union of two sets)



superduper.com



pharmax.com



bluetotal.com



mycanadashop.com

Organizing the sites

Content clustering applied to Web content

- Lots of different URLs/domains point to similar/identical sites
Why?
- Cluster on Q-gram similarity of DOM (i.e., HTML) content – 4-byte strings
Jaccard index (basically intersection/union of two sets)

Category tagging applied to clusters

- Three “categories”: **pharmaceuticals**, **replica** products and **software**
- Ad hoc keywords (brand names: e.g., Viagra, Rolex, generic terms: watches, herbal, etc). Why these?

Program tagging applied to tagged clusters

- Web content signatures created to match to individual affiliate programs

Many “brands” for same affiliate program



Output of clustering

<i>Stage</i>	<i>Pharmacy</i>	<i>Software</i>	<i>Replicas</i>	<i>Total</i>
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

Table III: Breakdown of clustering and tagging results.

Note: ~69k domains – out of 3M unique domains crawled

- What might concern be?

Opportunities for Bias/Error?

- Content clustering
- Category tagging
- How does paper validate?
- Years later, automation:
Knock It Off: Profiling the Online Storefronts of Counterfeit Merchandise, KDD 2014

Where do program tag signatures come from?

Even with our conservative approach, a concern is that our keyword matching heuristics might have missed a site of interest. Thus, for the remaining untagged clusters, we manually checked for such false negatives, i.e., whether there were clusters of storefront pages selling one of the three goods that should have a category tag, but did not. We examined the pages in the largest 675 untagged clusters (in terms of number of pages) as well as 1,000 randomly selected untagged clusters, which together correspond to 39% of the URLs we crawled. We did not find any clusters with storefronts that we missed.⁶

Affiliate Program	URLs	Volume	Domains
RX-Promotion	160,522,026	21.7%	10,586
Mailien	69,961,211	23.57%	14,444
Pharmacy Express	69,959,633	23.57%	14,381
ED Express	1,578	<0.01%	63
ZedCash (Pharma)	42,297,130	18.93%	6,981
Dr. Maxman	32,184,860	13.19%	5,641
Viagrow	5,222,658	3.57%	386
US HealthCare Inc.	3,196,538	1.42%	167
MaxGentleman	1,144,703	0.39%	672
VigREX	426,873	0.31%	39
Stud Extreme	71,104	0.05%	43
ManXtenz	50,394	<0.01%	33
GlavMed	28,313,136	7.84%	2,933
Online Pharmacy	17,266,034	5.07%	2,922
EvaPharmacy	12,798,999	7.91%	11,285
World Pharmacy	10,412,850	5.88%	691
PH Online	2,971,368	2.14%	101
Swiss Apotheke	1,593,532	0.21%	118
HerbalGrowth	265,131	0.19%	17
RX Partners	229,248	0.15%	448
Stimul-cash	157,537	0.07%	50
MAXX Extend	104,201	<0.01%	23
DrugRevenue	51,637	0.05%	122
Ultimate Pharmacy	44,126	0.02%	12
Greenline	25,021	<0.01%	1,766
Virility	23,528	0.01%	9
MediTrust	6,156	<0.01%	24
RX Rev Share	5,690	<0.01%	183
Unknown Program	3,310	<0.01%	1,270
Canadian Pharmacy	1,392	<0.01%	133
RXCash	287	<0.01%	22
Stallion	80	<0.01%	2
Pharma Total	347,053,630	93.74%	54,142

Affiliate Program	URLs	Volume	Domains
Royal Software	2,291,571	1.48%	572
EuroSoft	694,810	0.31%	1,161
Auth. Soft. Resellers	65,918	<0.01%	4,117
OEM Soft Store	19,436	<0.01%	1,367
Soft Sales	93	<0.01%	35
Software Total	3,071,828	1.79%	7,252

Another take on clustering

What is purpose of this table?

Affiliate Program	URLs	Volume	Domains
ZedCash (Replica)	13,264,108	4.29%	7,011
Ultimate Replica	10,464,930	3.35%	5,032
Distinction Replica	1,252,816	0.3%	130
Diamond Replicas	506,486	0.14%	1,307
Prestige Replicas	382,964	0.16%	101
Exquisite Replicas	620,642	0.32%	128
One Replica	21,318	0.02%	83
Luxury Replica	11,207	<0.01%	28
Aff. Accessories	3,669	<0.01%	187
Swiss Rep. & Co.	76	<0.01%	15
WatchShop	2,086,930	0.17%	547
Replica Total	15,351,038	4.46%	7,558

Last part: purchasing

Why purchase?

- **Payment**

Are there extra charges on card (is it a scam?)

Who took the money (another part of infrastructure)

- **Fulfillment**

Do you get items at all? Are they real? (insight into business)

Where did the product ship from (another part of infrastructure)

Customer service (another part of infrastructure)

Quick review: how an open-loop card transaction works...

Buyer

SECURE ORDER FORM
100% SECURE CREDIT CARD TRANSACTION

WE ACCEPT


Order details

Name	Price	Quantity	Sum
Zyrtec (Generic) 60 pills x 5 mg	\$51.73	1	\$51.73
Delivery type Courier			\$14.95
	Total		\$66.68



For your safety we use highly secure order processing server with our own secure certificate

[return to shop](#)

Billing Address

First Name

Last name

Shipping Address

☒ Shipping info equals to Billing Info



Supplier

Supplier Completes Transaction
Based on Authorization Response



Issuer

What's in an authorization request? (ie, from merchant's bank to your issuer's bank)

User info

- Name, card number, expiration, CVV
- Address components (for AVS)

Amount

Acquirer info

- **Bank Identification Number (BIN)** – six digits identify acquiring bank
- **Merchant descriptor** – the text string of merchant (arbitrary)
- **Card Acceptor ID** – terminal/account identifier at bank
- **Merchant Category Code (MCC)** – kind of items sold (e.g., 5912 drugs/pharmacy)

What does a credit card authorization record look like?

CANADIAN
Health&Care Mall

ALL PRODUCTS | ABOUT US | HOW TO ORDER | TESTIMONIALS | FAQ | CONTACTS

Bank Identification Number (BIN) 448314



Merchant descriptor “Smart rt online”

Card Acceptor ID “8875236”

Merchant Category Code (MCC) 5912
Drug Stores & Pharmacies

Propecia ★
Maxaman ★
Super Active ED Pack

[More Info](#)

What's in an authorization request? (ie, from merchant's bank to your issuer's bank)

User info

- Name, card number, expiration, CVV
- Address components (for AVS)

Amount

Acquirer info

- **Bank Identification Number (BIN)** – six digits identify acquiring bank
- **Merchant descriptor** – the text string of merchant (arbitrary)
- **Card Acceptor ID** – terminal/account identifier at bank
- **Merchant Category Code (MCC)** – kind of items sold (e.g., 5912 drugs/pharmacy)

Bottom line: issuing bank **MUST** know who is getting money

- But you need to be a bank...

Side story – how does UCSD become the equivalent of a bank?

First try: Visa gift cards

- Idea: when you call Visa gift card company potentially contest a card (i.e., chargeback) they give you an Acquirer Reference Number (ARN) which is the same as the merchant bank BIN
 - Scales poorly – calls with customer service; suspicious
 - Find gift card with online ARN portal
- Update to PATRIOT act makes gift cards (at the time) useless for international transactions

Chris Kanich (then Ph.D. student, now faculty at UIC) cold calls banks/specialty issuers until he finds one willing to help us

- We ultimately “rent” a BIN (i.e., first six digits of credit card #)
- Get “bank view” of transactions

Aside: even with this, purchasing is *hard*

Operational issues

- Geo-located IPs matched with Visa cards
 - In some site's fraud checks
- Distinct credit card numbers & phone #'s, email
 - Unsolicited customer service calls: need to "stay in character"
- Where to receive goods?

Institutional issues

- Making purchases from criminals? Liability?
- Audit issues on cash-equivalent transactions
- Expect to have money **stolen**
- What to do with received counterfeit goods?

















600+ orders later...



Special thanks to the Office of the VCR, GC and UCOP Office of Research Compliance

“Intervention analysis”

Look for impact of removing one or more resources in:

- Registrar

- DNS hosting

- Web hosting

- Payments

(i.e., are there economic “bottlenecks” at any of these tiers?)

What is the denominator in such calculations?

(i.e., bottleneck in “what”)?

Ok, so blowing up hosting provider X will block 35% of spam... what else might we care about?

“Intervention analysis”

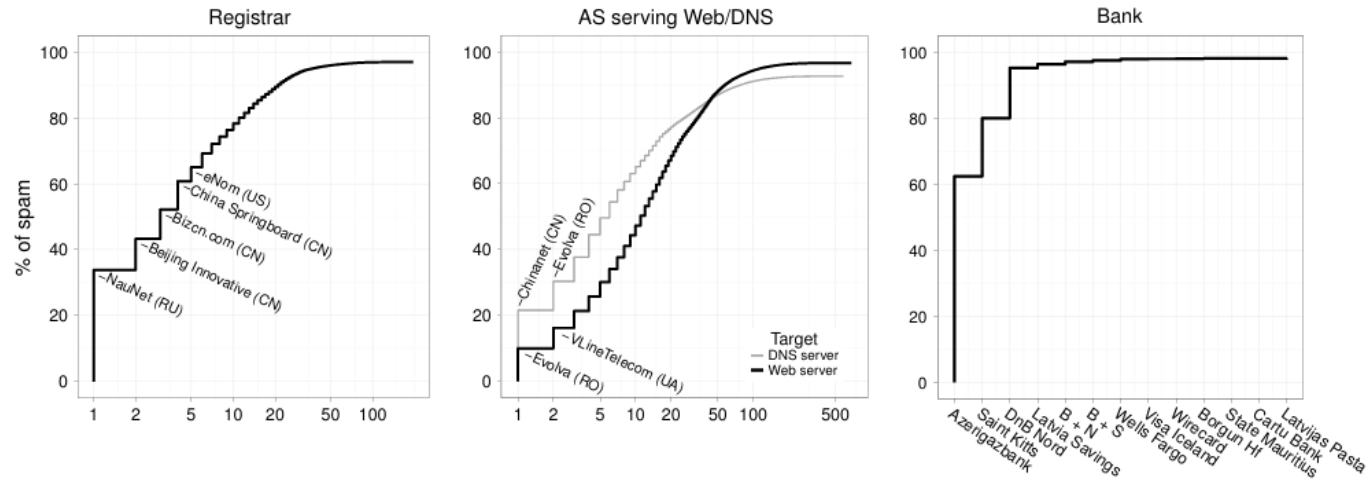
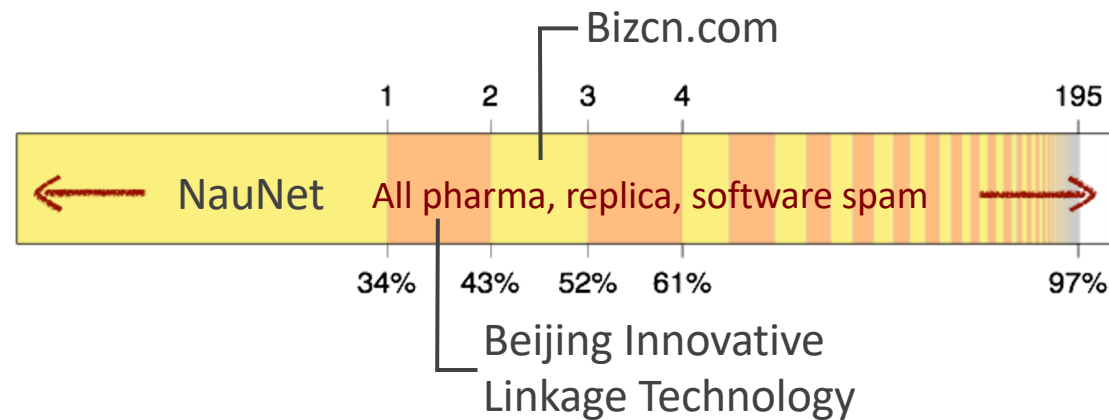


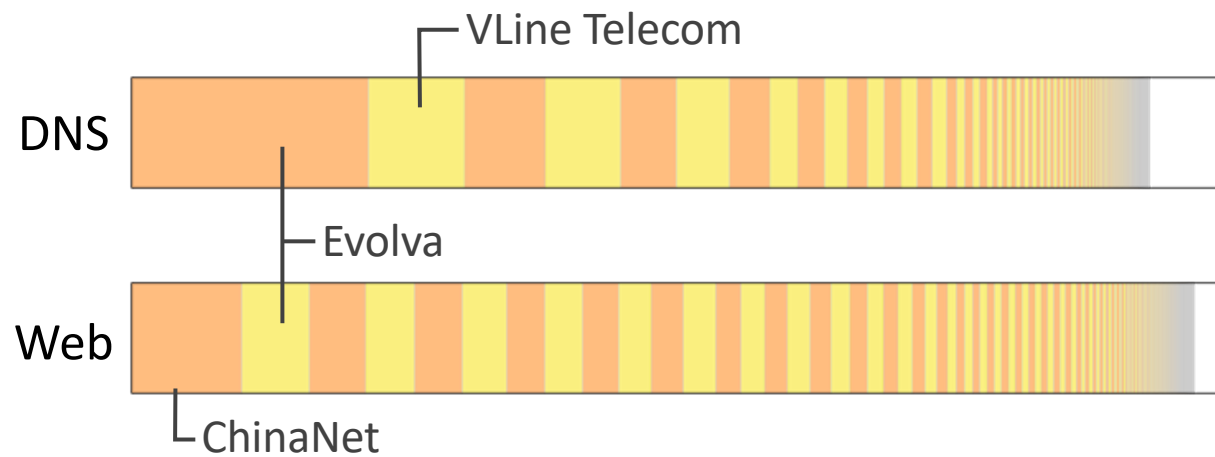
Figure 5: Takedown effectiveness when considering domain registrars (left), DNS and Web hosters (center) and acquiring banks (right).

Registrars



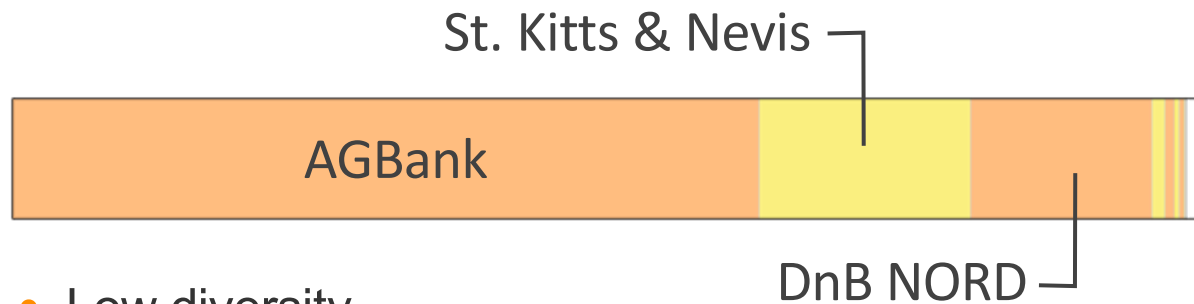
- Some concentration in NauNet (Russia), BILT, Bizcn (China)
 - Lots of diversity in remaining domains
- But many alternatives, low switching cost; hard to have lasting impact (see Liu et al, LEET 2011)

DNS and Web hosting



- Even more diversity than with registrars
- Again many alternatives, low switching cost

Merchant banks



- Low diversity
 - 3 banks covered 95% of spam
 - Fewer banks willing handle “high-risk” merchants
- High switching cost
 - In-person account creation, due diligence, multi-day process
 - Upfront capital, holdback forfeiture

Paper proposes following interventions

Takedown: acquiring (merchant) bank

- Pressure banks to drop customers
- Challenges: bi-lateral process, potentially slow, incentives not aligned

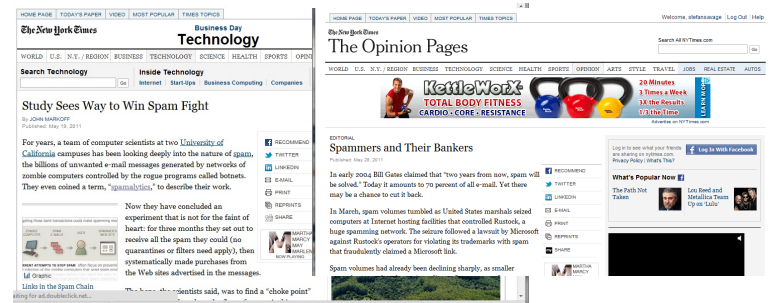
Blocklist: issuing (consumer) banks

- Few banks issue majority of US Visa/MC
- Could demonetize spam ecosystem by refusing certain transactions with “bad” acquirers; fast
- Challenges: incentives not aligned

Thoughts?

Epilogue: transferring results

Major press



Govt briefings

- FDA, FTC, DHS & EOP, Commerce parts of intel community
- Support for active LE investigations



Industry briefings

- Banks, investigations, pharma



Key insight: banks don't care, brands care

[ROGUEBLOCK INITIATIVE](#)[ROGUEBLOCK MEMBER LOGIN](#)[REQUEST MORE INFO](#)

IACC PAYMENT PROCESSOR INITIATIVE (RogueBlock®)



IACC RogueBlock® Initiative

The Payment Processor Initiative (RogueBlock®) is a collaborative effort of the IACC and the payment industry to create a streamlined, simplified procedure for members to report online sellers of counterfeit or pirated goods directly to credit card and payment processing networks.

With a goal of facilitating prompt action against counterfeiters' merchant accounts and diminishing the ability of such sellers to profit from their illicit sales, the Payment Processor Initiative (RogueBlock®)

Result: targeted payment intervention regime

Undercover test purchase at counterfeit site

- Get merchant bank BIN from transaction

IP holder notifies card network (e.g., Visa/MC)

- Investigation; complaint delivered to merchant bank

Leverage existing card association contract with bank

- Products must be legal in selling country *and purchasing country*
- Merchant bank owns liability
- Fines, increased scrutiny, threat of de-association

Merchant account shutdown

- Bank holdback ~10% revenues over 180 days for high-risk customers
Against future liabilities
- Empirically, money is just taken when account is shut down for cause

Priceless paper: weird paper

We had helped create and were helping to run the operational takedown program

- Test case with Microsoft, then Pfizer
- This paper was to evaluate how well that worked
- But was not public that this was going on; so odd phrasings in paper
"By opportunistically combining our own active purchasing data with contemporary disruption efforts by brand-holders and payment card networks, we gather the first empirical dataset concerning this approach."

What was quantitative measurement technique?

- Buy stuff
From affiliate programs before and after takedown request from brand holder
- What do you learn from buying stuff (i.e., what is signal?)
Were you able to buy stuff? Or was payment declined?
If you were successful, what bank took your money?

There was also a qualitative measurement technique?

- Join affiliate programs (i.e., pretending to be criminal) and see what internal messaging is during takedowns

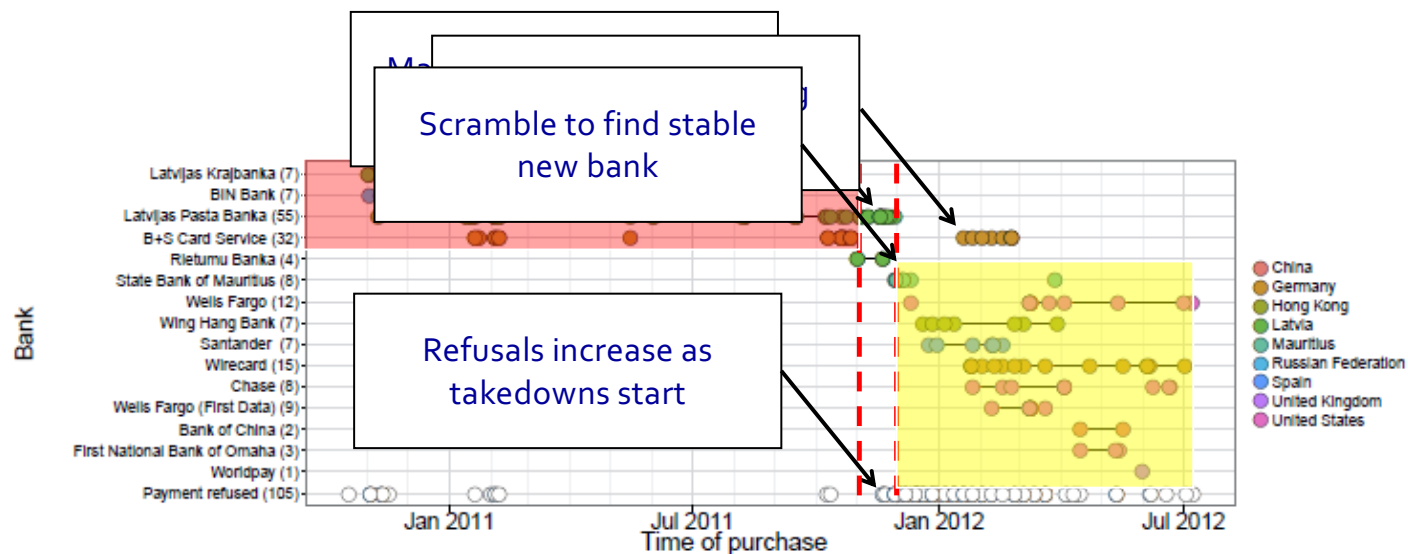
OEM (pirate) software affiliate programs (circa 2012)



OEM software story

Microsoft Thanksgiving surprise (Nov '11)

- Methodically issued complaints for accounts of **every** software affiliate program
- Diligent follow up: new processing -> new complaints



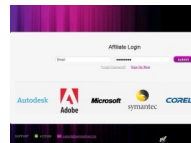
Qualitative Timeline

11/2011: Undercover purchases and referrals to Visa/MC start

11/20/2011: ATTENTION! Deactivated servers, y having problems with the problem of accounts and accounts. We have decided to temporarily stop accepting OEM traffic.



2011-11-22 10:16:38 Starting today our work has stopped banking. Due to this we have a (and more) in this case it is an affiliate program for the distribution of shares and for processing.



1/23/2012 Remark by leading affiliate:
"The sun is setting on the OEM era"

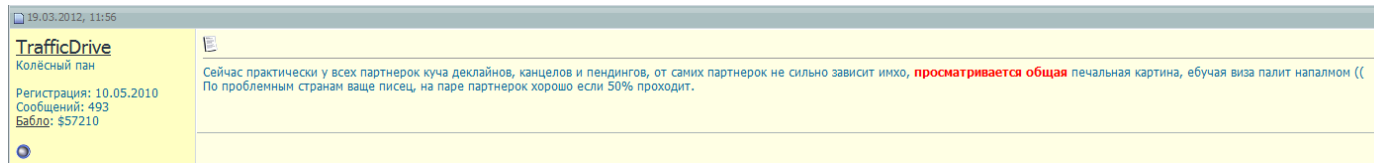
Similar story with pharmaceuticals: Affiliate programs operating in 2011



Similar story with pharmaceuticals Affiliate programs operating in 2013



Summary from a pharma affiliate program manger...



“Right now most affiliate programs have a mass of declines, cancels and pendings, and it doesn't depend much on the program imho, there is a general sad picture, **fucking Visa is burning us with napalm** (for problematic countries, it's totally fucked, on a couple of programs you're lucky if you get 50% through).”

Priceless study outcomes

As of mid-late 2013

- OEM software market was **decimated**
 - 90% of affiliate programs folded
 - Remaining programs stopped selling Microsoft software
 - MS lacked standing to issue complaints about other software (e.g., Adobe)
 - Single operation effective for 18 months
- Similar story in pharma market
 - 50% of organizations close
- European banks depart “risky” markets
- But... lots of adaptations (note this is all pre crypto currency growth)

Instrumental in helping understand importance of aligning incentives with measurement framing

Next class

Neha Chachra, Damon McCoy, Stefan Savage, Geoffrey M. Voelker. [Empirically Characterizing Domain Abuse and the Revenue Impact of Blacklisting](#), In *Proceedings of the Workshop on the Economics of Information Security (WEIS)*, June 2014, pp. 4:1:1–4:1:13