Survival analysis techniques for studying cybercrime

Tyler Moore

Computer Science & Engineering Department, SMU, Dallas, TX

November 1, 2012

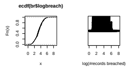
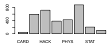
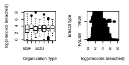
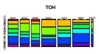
Notes

Outline

1. Case-control studies for analyzing data
   - Case study: Spear-phishing study
   - Case study: Search-redirection attacks
2. Survival analysis
   - Definitions
   - Case study: Phishing website recompromise

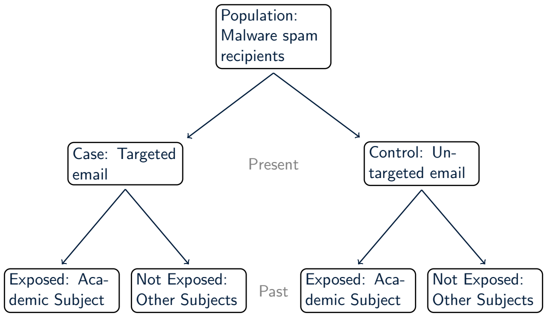
Notes

Guide to analyzing data

Type of Data	Exploration	Statistics	RByEx
1 numerical variable		one way t-test, Wilcox test	6.3
1 categorical variable		–	3.1
# categories=2	–	prop.test	6.2
1 categorical, 1 numerical		anova, Permutation	10
# categories=2	–	2-way t, Wilcox test, Perm.	6.4
2 categorical variables		χ^2 test	3.2–3.5

Notes

Case-control study: spear phishing and academic specialty



Population: Malware spam recipients

Case: Targeted email

Present

Control: Untargeted email

Exposed: Academic Subject

Not Exposed: Other Subjects

Past

Exposed: Academic Subject

Not Exposed: Other Subjects

Paper available for download in Blackboard: “Who’s next? Identifying risk factors for subjects of targeted attacks”

Notes

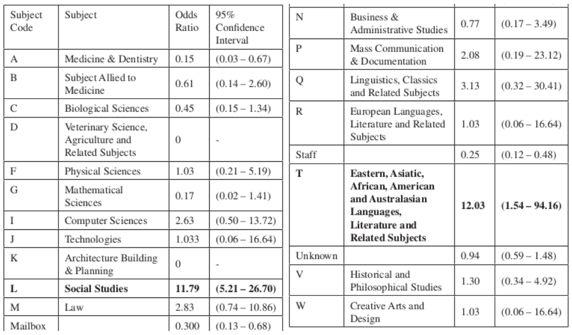
The odds ratio

	Case (afflicted)	Control (not afflicted)
Exposed (has risk factor)	p_{11}	p_{10}
Not exposed (no risk factor)	p_{01}	p_{00}

$$\text{odd's ratio} = \frac{p_{11} * p_{00}}{p_{10} * p_{01}}$$

Notes

Odds ratios for academic subjects in spear phishing study



Subject Code	Subject	Odds Ratio	95% Confidence Interval
A	Medicine & Dentistry	0.15	(0.03 – 0.67)
B	Subject Allied to Medicine	0.61	(0.14 – 2.60)
C	Biological Sciences	0.45	(0.15 – 1.34)
D	Veterinary Science, Agriculture and Related Subjects	0	-
F	Physical Sciences	1.03	(0.21 – 5.19)
G	Mathematical Sciences	0.17	(0.02 – 1.41)
I	Computer Sciences	2.63	(0.50 – 13.72)
J	Technologies	1.033	(0.06 – 16.64)
K	Architecture Building & Planning	0	-
L	**Social Studies**	**11.79**	**(5.21 – 26.70)**
M	Law	2.83	(0.74 – 10.86)
Mailbox		0.300	(0.13 – 0.68)
N	Business & Administrative Studies	0.77	(0.17 – 3.49)
P	Mass Communication & Documentation	2.08	(0.19 – 23.12)
Q	Linguistics, Classics and Related Subjects	3.13	(0.32 – 30.41)
R	European Languages, Literature and Related Subjects	1.03	(0.06 – 16.64)
Staff		0.25	(0.12 – 0.48)
T	**Eastern, Asiatic, African, American and Australasian Languages, Literature and Related Subjects**	**12.03**	**(1.54 – 94.16)**
Unknown		0.94	(0.59 – 1.48)
V	Historical and Philosophical Studies	1.30	(0.34 – 4.92)
W	Creative Arts and Design	1.03	(0.06 – 16.64)

Notes

Illicit online pharmacies

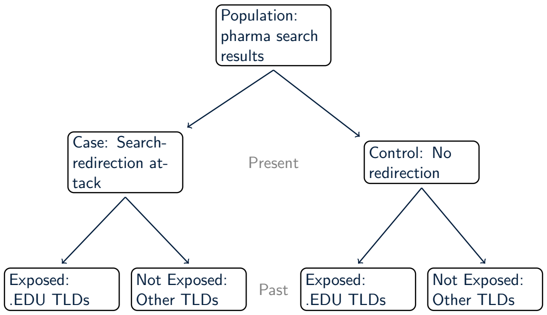
Notes

Illicit online pharmacies

- What do illicit online pharmacies have to do with phishing?
- Both make use of a similar criminal supply chain
  1. **Traffic**: hijack web search results (or send email spam)
  2. **Host**: compromise a high-ranking server to redirect to pharmacy
  3. **Hook**: affiliate programs let criminals set up website front-ends to sell drugs
  4. **Monetize**: sell drugs ordered by consumers
  5. **Cash out**: no need to hire mules, just take credit cards!
- For more: `http://lyle.smu.edu/~tylerm/usenix11.pdf`

Notes

Case-control study: search-redirection attacks



Notes

Case-control study: search-redirection attacks

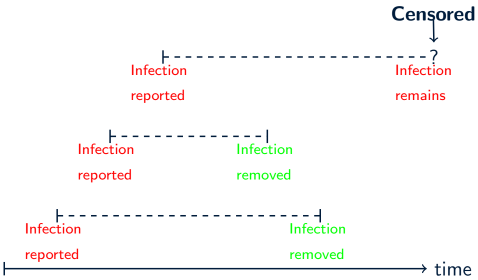
R code: `http://lyle.smu.edu/~tylerm/courses/econsec/code/pharmaOdds.R`

Data format:

Date	Search Engine	Search Term	Pos.	URL	Domain	Redirects?	TLD
2011-11-03	Google	20 mg ambien overdose	1	http://products.sanofi.us/ambien/ambien.pdf	sanofi.us	False	other
2011-11-03	Google	20 mg ambien overdose	2	http://swift.sonoma.edu/education/newton/newtonsLaws/?20-mg-ambien-overdose	sonoma.edu	False	.EDU
2011-11-03	Google	20 mg ambien overdose	3	http://ambienoverdose.org/about-2/	ambienoverdose.org	False	.ORG
2011-11-03	Google	20 mg ambien overdose	4	http://answers.yahoo.com/question/index?qid=20090712025803AA10g8Z	yahoo.com	False	.COM
2011-11-03	Google	20 mg ambien overdose	5	http://en.wikipedia.org/wiki/Zolpidem	wikipedia.org	False	.ORG
2011-11-03	Google	20 mg ambien overdose	6	http://blocsonic.com/blog	blocsonic.com	False	.COM
2011-11-03	Google	20 mg ambien overdose	7	http://dinarvets.com/forums/index.php?/user/39154-ambien-side-effects/page	dinarvets.com	False	.COM
2011-11-03	Google	20 mg ambien overdose	8	http://nemo.mwd.hartford.edu/mwd08/images/?20-mg-ambien-overdose	hartford.edu	True	.EDU
2011-11-03	Google	20 mg ambien overdose	9	http://www.formspring.me/AmbienCheapOn	formspring.me	False	other
2011-11-03	Google	20 mg ambien overdose	11	http://www.drugs.com/pro/zolpidem.html	drugs.com	False	.COM
2011-11-03	Google	20 mg ambien overdose	12	http://www.engineer.tamuk.edu/departments/ieen/images/ambien.html	tamuk.edu	False	.EDU
2011-11-03	Bing	20 mg ambien overdose	1	http://answers.yahoo.com/question/index?qid=20090712025803AA10g8Z	yahoo.com	False	.COM
2011-11-03	Bing	20 mg ambien overdose	2	http://www.healthcentral.com/sleep-disorders/h/20-mg-ambien-overdose.html	healthcentral.com	False	.COM
2011-11-03	Bing	20 mg ambien overdose	3	http://ambien20mg.com/	ambien20mg.com	False	.COM
2011-11-03	bing	20 mg ambien overdose	4	http://www.chacha.com/question/will-20-mg-of-ambien-cr-get-you-high	chacha.com	True	.COM
2011-11-03	bing	20 mg ambien overdose	5	http://www.rxlist.com/ambien-drug.htm	rxlist.com	True	.COM
2011-11-03	Bing	20 mg ambien overdose	6	http://www.drugs.com/pro/zolpidem.html	drugs.com	False	.COM
2011-11-03	Bing	20 mg ambien overdose	7	http://answers.yahoo.com/question/index?qid=20111024222432AARFvPB	yahoo.com	False	.COM
2011-11-03	Bing	20 mg ambien overdose	8	http://en.wikipedia.org/wiki/Zolpidem	wikipedia.org	False	.ORG
2011-11-03	Bing	20 mg ambien overdose	9	http://www.thefullwiki.org/Sertraline	thefullwiki.org	False	.ORG
2011-11-03	bing	20 mg ambien overdose	10	http://www.rxlist.com/edluar-drug.htm	rxlist.com	True	.COM
2011-11-03	Bing	20 mg ambien overdose	11	http://www.formspring.me/ambienpill	formspring.me	False	other
2011-11-03	Bing	20 mg ambien overdose	12	http://ambiendosage.net/	ambiendosage.net	False	.NET

Notes

Survival analysis)



Notes

Censored data happens a lot

- Real-world situations
  - Life-expectancy
  - Criminal recidivism rates
- Cybercrime applications
  - Measuring time to remove X (where X=malware, phishing, scam website, . . .)
  - Measuring time to compromise
  - Measuring time to re-infection
- Best resource I found on survival analysis in R: `http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf`

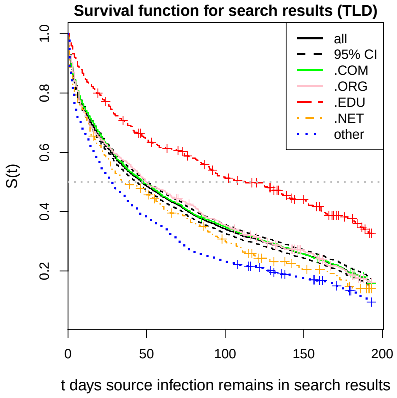
Notes

Survival analysis (package `survival` in R)

- Key challenge: estimating probability of survival when some data points survive at the end of the measurement
  - Solution: use the Kaplan-Meier estimator to compute probabilities that account for samples still alive (`survfit` in R)
- Common qeustion: Are survival functions split over categorical variables statistically different
  - Use the log-rank test (`survfit` in R)
  - Analagous to χ^2 test
- Cox-proportional hazard model is a more sophisticated way to see how multiple variables affect the *hazard rate*
  - Hazard function $h(t)$: expected number of failures during the time period t

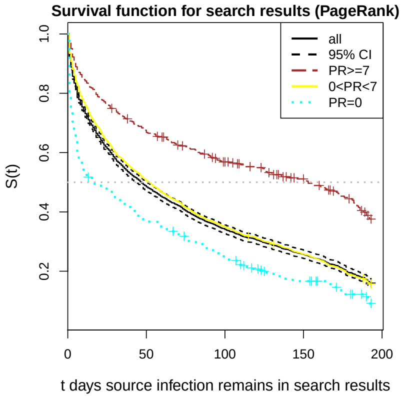
Notes

Pharmacy redirection duration by TLD



Notes

Pharmacy redirection duration by PageRank



Notes

Statistics disentangle effect of TLD, PageRank on duration

Cox-proportional hazard model
$h(t) = \exp(\alpha + \text{PageRank}x_1 + \text{TLD}x_2)$

	coef.	exp(coef.)	Std. Err.)	Significance
PageRank	-0.079	0.92	0.0094	$p < 0.001$
`.edu`	-0.26	0.77	0.084	$p < 0.001$
`.net`	0.10	1.1	0.081	
`.org`	0.055	1.1	0.052	
other TLDs	0.34	1.4	0.053	$p < 0.001$

log-rank test: Q=159.6, $p < 0.001$

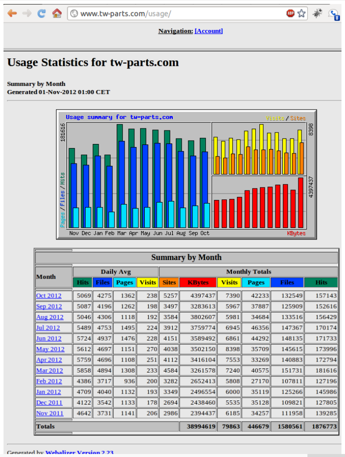
Notes

Phishing website recompromise

- Full paper: `http://lyle.smu.edu/~tylerm/cs81.pdf`
- What constitutes recompromise?
  - If one attacker loads two phishing websites on the same server a few hours apart, we classify it as one compromise
  - If the phishing pages are placed into different directories, it is more likely two distinct compromises
- For simplicity, we define website recompromise as distinct attacks on the same host occurring ≥ 7 days apart
- 83% of phishing websites with recompromises ≥ 7 days apart are placed in different directories on the server

Notes

The Webalizer



Navigation: [Account]

Usage Statistics for tw-parts.com

Summary by Month
Generated 01-Nov-2012 01:00 CET

Summary by Month

Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
Oct 2012	5069	4275	1362	238	5257	4397437	7390	42233	132549	157143
Sep 2012	5087	4196	1262	198	3497	3283613	5967	37887	125909	152616
Aug 2012	5046	4306	1118	192	3584	3802607	5981	34684	133516	156429
Jul 2012	5489	4753	1495	224	3912	3759774	6945	46356	147367	170174
Jun 2012	5724	4937	1476	228	4151	3589492	6861	44292	148135	171733
May 2012	5612	4697	1151	270	4038	3502150	8398	35709	145615	173996
Apr 2012	5759	4696	1108	251	4112	3416104	7553	33269	140883	172794
Mar 2012	5858	4894	1308	233	4584	3261578	7240	40575	151731	181616
Feb 2012	4386	3717	936	200	3282	2652413	5808	27170	107811	127196
Jan 2012	4709	4040	1132	193	3349	2496554	6000	35119	125266	145986
Dec 2011	4122	3542	1133	178	2694	2438460	5535	35128	109821	127805
Nov 2011	4642	3731	1141	206	2986	2394437	6185	34257	111958	139285
Totals						38994619	79863	446679	1580561	1876773

Generated by Webalizer Version 2.23

- Web page usage statistics are sometimes set up by default in a world-readable state
- We automatically checked all sites reported to our feeds for the Webalizer package, revealing over 2 486 sites from June 2007–March 2008
- 1 320 (53%) recorded search terms obtained from 'Referrer' header in the HTTP request
- Using these logs, we can determine whether a host used for phishing had been discovered using targeted search

Notes

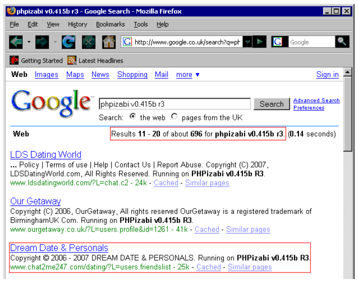
Types of evil search

- Vulnerability searches: `phpizabi v0.848b c1 hfp1` (unrestricted file upload vuln.), `inurl: com_juser` (arbitrary PHP execution vuln.)
- Compromise searches: `allintitle:  welcome paypal`
- Shell searches: `intitle: ''index of'' r57.php, c99shell drwxrwx`

Search type	Websites	Phrases	Visits
Any evil search	204	456	1 207
Vulnerability search	126	206	582
Compromise search	56	99	265
Shell search	47	151	360

Notes

One phishing website compromised using evil search



phpizabi v0.415b r3

Search: ● the web ○ pages from the UK

Web Results **11** - **20** of about **696** for **phpizabi v0.415b r3**. (**0.14** seconds)

LDS Dating World
... Policy | Terms of use | Help | Contact Us | Report Abuse. Copyright (C) 2007, LDSDatingWorld.com, All Rights Reserved. Running on **PHPizabi v0.415b R3**.
www.ldsdatingworld.com/?L=chat.c2 - 24k - Cached - Similar pages

Our Getaway
Copyright (C) 2006, OurGetaway, All rights reserved OurGetaway is a registered trademark of BirminghamUK Com. Running on **PHPizabi v0.415b R3**.
www.ourgetaway.co.uk/?L=users.profile&id=1261 - 41k - Cached - Similar pages

Dream Date & Personals
Copyright © 2006 - 2007 DREAM DATE & PERSONALS. Running on **PHPizabi v0.415b R3**.
www.chat2me247.com/dating/?L=users.friendslist - 25k - Cached - Similar pages

Notes

One phishing website compromised using evil search

```
1: 2007-11-30 10:31:33 phishing URL reported: http://chat2me247.com
/stat/q-mono/pro/www.lloydstsb.co.uk/lloyds_tsb/logon.ibc.html
2: 2007-11-30      no evil search term                  0 hits
3: 2007-12-01      no evil search term                  0 hits
4: 2007-12-02      phpizabi v0.415b r3                   1 hit
5: 2007-12-03      phpizabi v0.415b r3                   1 hit
6: 2007-12-04 21:14:06 phishing URL reported: http://chat2me247.com
/seasalter/www.usbank.com/online_banking/index.html
7: 2007-12-04      phpizabi v0.415b r3                   1 hit
```

Notes

Let's work with the data

R code: `http://lyle.smu.edu/~tylerm/courses/econsec/code/surviveEvil.R`

Data format:

TLD	1st Compromise	2nd Compromise	# days	Censored	Evil searches?
com	2008-01-28	2008-03-31	63	0	TRUE
com	2007-11-23	2008-03-31	129	0	TRUE
IP	2008-01-16	2008-03-31	75	0	TRUE
com	2008-01-16	2008-03-31	75	0	TRUE
com	2007-10-28	2007-11-06	8	1	TRUE
com	2008-01-20	2008-03-31	71	0	TRUE
jp	2007-11-12	2008-03-31	140	0	TRUE
nu	2008-01-31	2008-03-31	60	0	TRUE
net	2007-12-27	2008-03-31	95	0	TRUE
com	2008-02-08	2008-03-31	52	0	TRUE
IP	2007-12-07	2008-01-07	31	1	TRUE
IP	2008-01-29	2008-03-31	62	0	TRUE
com	2007-10-22	2007-11-14	22	1	TRUE
com	2008-01-22	2008-03-31	69	0	TRUE

Notes

Notes

Notes