EMBER: A Global Perspective on Extreme Malicious Behavior

Tamara Yu

Richard Lippmann, James Riordan, Stephen Boyer

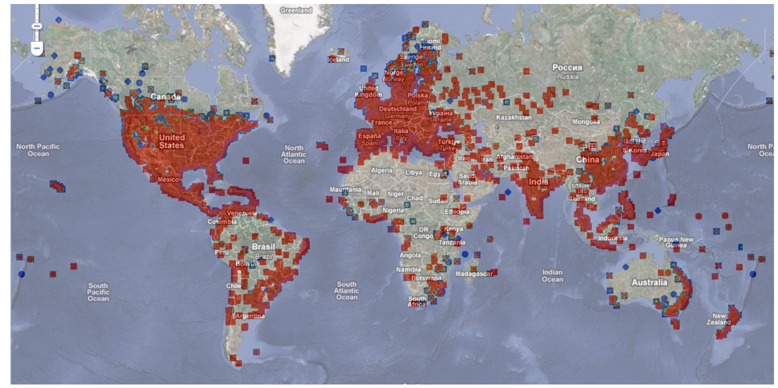
14 September 2010

MIT Lincoln Laboratory

This work is sponsored by the United States Air Force under Air Force Contract FA8721-05-C-0002 and by the National Security Agency/Central Security Service (Publication No. U/OO/NTC/0701). Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

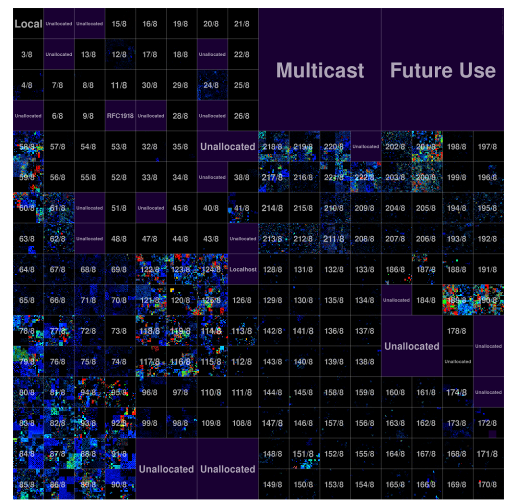
World Map for Security Visualization

- **World maps are commonly used for visualizing wide-spread malicious behavior of Internet hosts**
  - Pro: easy to understand
  - Con: generally not very useful



Conficker World Infections[1]

- **Recent security visualization research focuses on network-oriented views**
  - Cyber neighborhoods are deemed more relevant for threat analysis

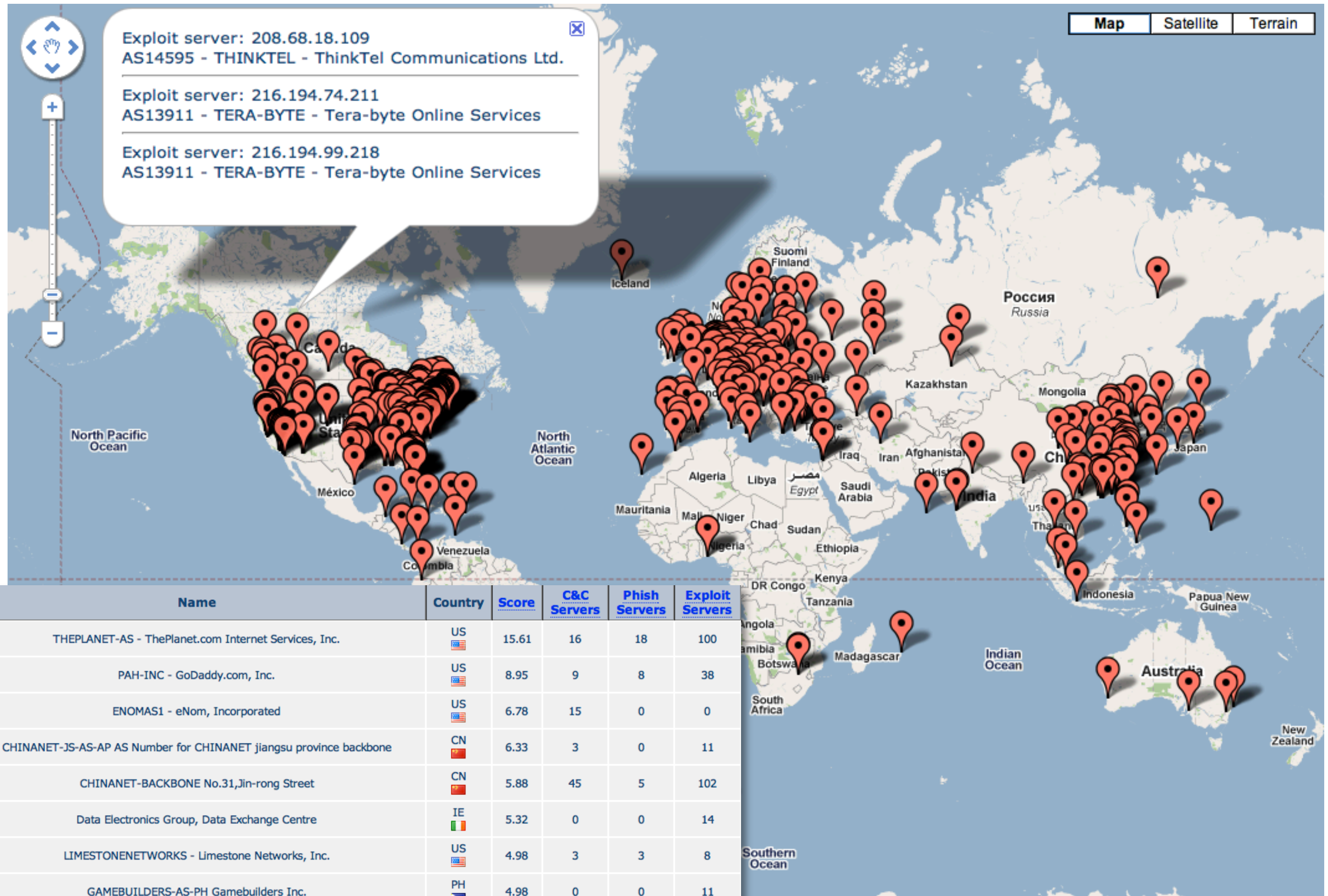


Conficker Network Neighborhood Infection Map[1]

Has the world map been all but written off as a "serious" security visualization?

[1] Conficker Working Group. Infection Maps, 2009. http://www.confickerworkinggroup.org/wiki/pmwiki.php/ANY/InfectionDistribution

Exhibit A: Dots on the Map



FIRE Global Map of Rogue Servers[1]

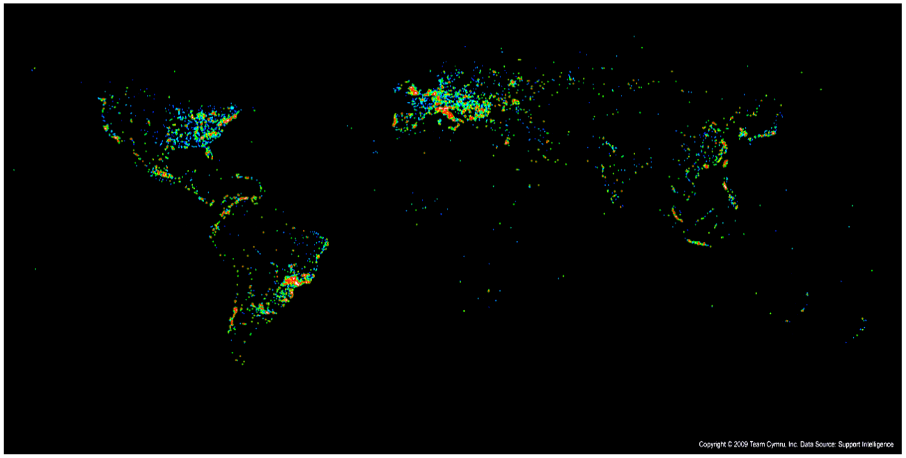
Rank	Rank Change	ASN	Name	Country	Score	C&C Servers	Phish Servers	Exploit Servers
1	✕	AS21844	THEPLANET-AS - ThePlanet.com Internet Services, Inc.	US	15.61	16	18	100
2	✕	AS26496	PAH-INC - GoDaddy.com, Inc.	US	8.95	9	8	38
3	↑	AS21740	ENOMAS1 - eNom, Incorporated	US	6.78	15	0	0
4	↑	AS23650	CHINANET-JS-AS-AP AS Number for CHINANET jiangsu province backbone	CN	6.33	3	0	11
5	↑	AS4134	CHINANET-BACKBONE No.31,Jin-rong Street	CN	5.88	45	5	102
6	↑	AS13100	Data Electronics Group, Data Exchange Centre	IE	5.32	0	0	14
7	↓	AS46475	LIMESTONENETWORKS - Limestone Networks, Inc.	US	4.98	3	3	8
8	↑	AS18018	GAMEBUILDERS-AS-PH Gamebuilders Inc.	PH	4.98	0	0	11
9	↓	AS24940	HETZNER-AS Hetzner Online AG RZ	DE	[illegible]	[illegible]	16	15
10	↓	AS15244	ADDD2NET-COM-INC-DBA-LUNARPAGES - Lunar Pages	US	4.58	3	5	9

Top Malicious Autonomous Systems

[1] FIRE: FInding RoguE Networks, 2010. http://maliciousnetworks.org/map.php

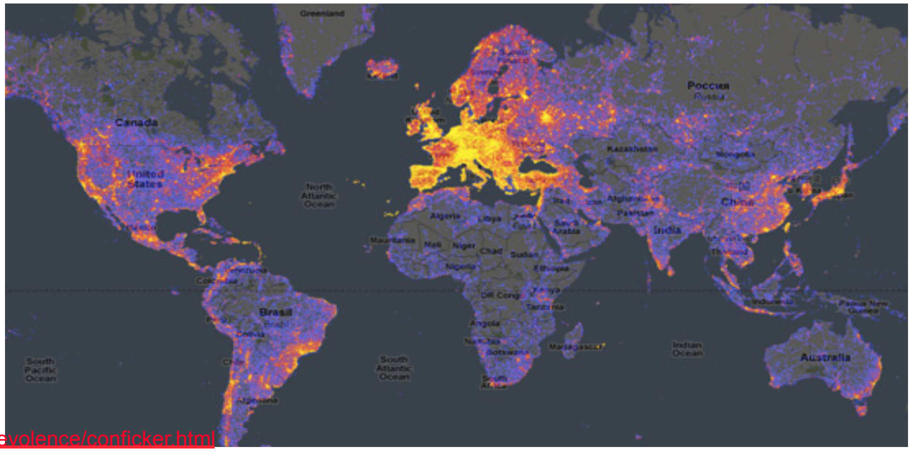
Exhibit B: Heat Maps

Conficker[1]



NASA "Earth-at-Night"[2]

"Touristiness"[3]

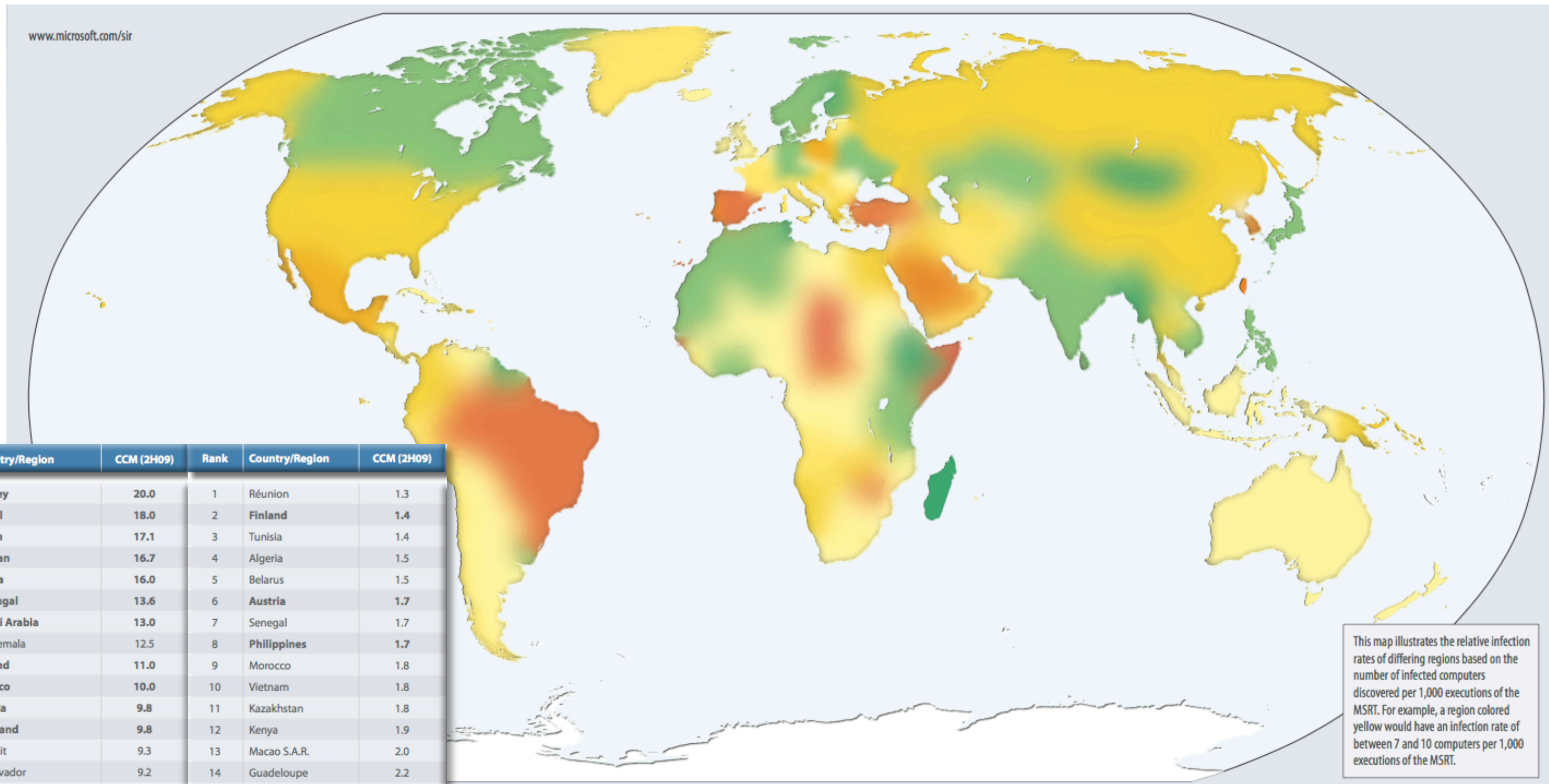


Heat map displays mainly show population centers, where most potential victims are…

…in the same way artificial lights or tourists show up in large cities.

[1] Team Cymru. Conficker Worm Visualizations, 2009. http://www.team-cymru.org/Monitoring/Malevolence/conficker.html
[2] NASA. Earth's City Lights, 2000. http://visibleearth.nasa.gov/view_rec.php?id=1438
[3] World Touristiness Map, 2010. http://www.bluemoon.ee/~ahti/touristiness-map/

Exhibit C: Normalized Heat Map



Rank	Country/Region	CCM (2H09)	Rank	Country/Region	CCM (2H09)
1	**Turkey**	**20.0**	1	Réunion	1.3
2	**Brazil**	**18.0**	2	**Finland**	**1.4**
3	**Spain**	**17.1**	3	Tunisia	1.4
4	**Taiwan**	**16.7**	4	Algeria	1.5
5	**Korea**	**16.0**	5	Belarus	1.5
6	**Portugal**	**13.6**	6	**Austria**	**1.7**
7	**Saudi Arabia**	**13.0**	7	Senegal	1.7
8	Guatemala	12.5	8	**Philippines**	**1.7**
9	**Poland**	**11.0**	9	Morocco	1.8
10	**Mexico**	**10.0**	10	Vietnam	1.8
11	**Russia**	**9.8**	11	Kazakhstan	1.8
12	**Thailand**	**9.8**	12	Kenya	1.9
13	Kuwait	9.3	13	Macao S.A.R.	2.0
14	El Salvador	9.2	14	Guadeloupe	2.2
15	Honduras	9.1	15	Pakistan	2.2
16	Colombia	9.1	16	**Germany**	**2.2**
17	**Hungary**	**8.9**	17	Jamaica	2.3
18	Croatia	8.9	18	**Switzerland**	**2.3**
19	Egypt	8.6	19	**Japan**	**2.3**
20	Serbia and Montenegro*	8.1	20	Puerto Rico	2.4
21	[illegible]	7.8	21	[illegible]	2.5
22	**Greece**	**7.7**	22	**Denmark**	**2.5**
23	Ecuador	7.6	23	Canada	2.5
24	**Israel**	**7.3**	24	**India**	**2.6**
25	**China**	**7.0**	25	Uruguay	2.6

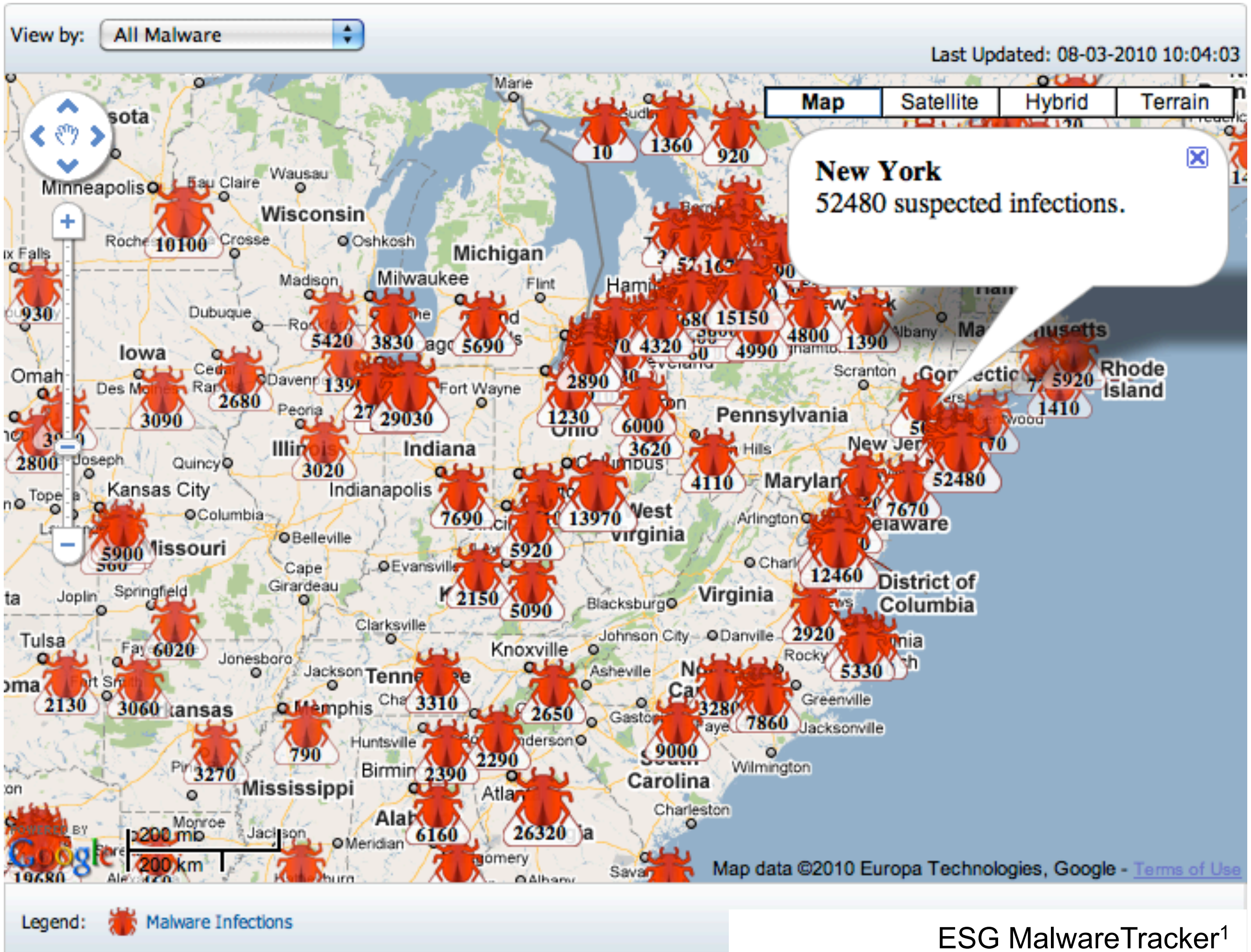
High-CCM Countries

Low-CCM Countries

Infection rates around the world, expressed in CCM, or number of computers cleaned per thousand executions of Malicious Software Removal Tool[1]

[1] Microsoft. Microsoft Security Intelligence Report Volume 8, May 2010. http://www.microsoft.com/downloads/details.aspx?FamilyID=2c4938a0-4d64-4c65-b951-754f4d1af0b5

Exhibit D: Infections by City



ESG MalwareTracker[1]

[1] Enigma Software Group. ESG MalwareTracker, 2010. http://www.enigmasoftware.com/malwaretracker/

Find Regions with Malicious Activity that is Higher or Lower than Expected

- **Group IP addresses by City**
  - Using countries is often too coarse
  - Internet service provider boundaries often agree with city boundaries
  - Internet security authorities and policies often apply across a city
  - Law enforcement domains often agree with city boundaries
  - Malware often preferentially spreads to local class C networks and these are often within a city
  - This granularity will make it possible to see targeted malware
- **Map IP addresses exhibiting malicious activity geographically to cities**
- **Normalize by the population of computers in each city**

Utility of Providing Plots of Extreme Variations In Malicious Activity

- **High Malicious Activity**
  - ISPs explicitly allow and protect criminal activity (e.g. the Russian Business Network)
  - Poor “network hygiene”
  - More highly targeted than other regions
- **Low Malicious Activity**
  - ISPs actively prevent, block or rapidly detect and eliminate malicious activity
  - Strong cyber laws and enforcement
  - Good “network hygiene”
  - Not being targeted by cyber criminals

Geo-Locate IP Addresses

- **Accuracy of the analysis is influenced by**
  - How malicious IP addresses are harvested
  - Geo-location accuracy

- **For proof-of-concept demonstration, we use**
  - MaxMind GeoLite City[1]: database for geo-locating IP addresses to cities
  - Dshield[2]: dataset of malicious IP addresses (approx. 600,000 daily)

# source IP	targetport	protocol	reports	targets	firstseen	lastseen
216.113.038.035	1080	6	147601	84012	6:46:07	22:43:31
088.084.131.145	22	6	143515	79580	2:58:26	16:32:07
094.023.193.116	8080	6	76089	76080	16:31:45	20:20:52
222.073.204.093	1433	6	66190	64490	0:12:52	22:01:51
200.020.215.131	22	6	119222	64348	7:29:59	7:43:38
061.160.213.136	2967	6	62741	62494	0:12:41	23:08:34
061.160.213.016	135	6	77907	57514	0:00:48	23:07:10
220.184.013.088	2967	6	81908	57240	1:20:45	23:52:41
058.243.161.051	1434	17	54275	54226	0:00:02	23:59:59
202.101.180.165	1434	17	44066	44040	0:00:02	23:59:59
061.189.153.251	1434	17	37270	37244	0:00:00	23:59:59

[1] MaxMind GeoLite City, 2010. http://www.maxmind.com/app/geolitecity
[2] DShield, 2010. http://www.dshield.org

Estimate City Computer Population

- **It is impossible to directly count the number of Internet hosts in a city**

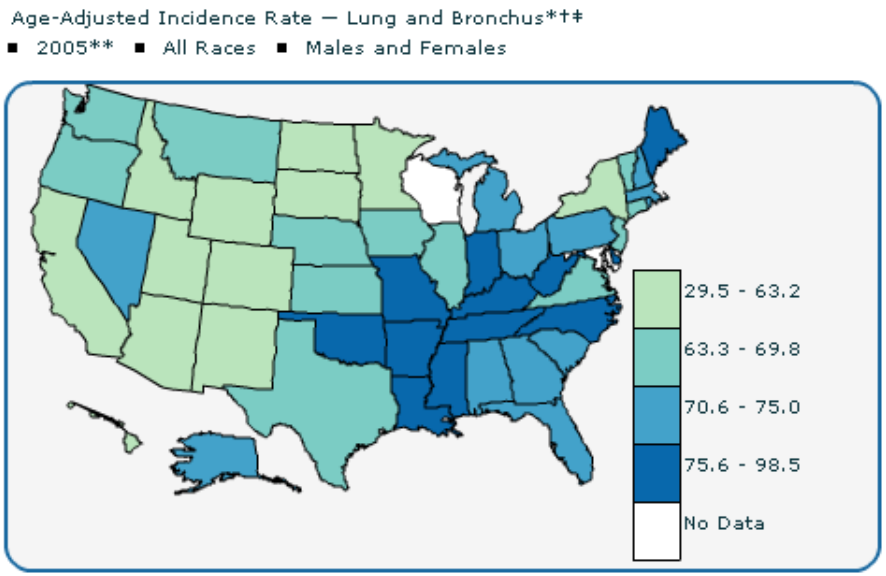
- **Approximation methods are either inaccurate or not scalable**
  - e.g., estimate from address allocation, active probing, or inference from web or DNS traffic

- **Our method relies on public data sources**
  - GeoNames[1]: city human population sizes
  - Internet World Stats[2]: country Internet penetration rates

$$n_{city} = \text{Population}_{city} \bullet \text{Internet Penetration Rate}_{city}$$

[1] GeoNames, 2010. http://www.geonames.org
[2] Internet World Stats, 2010. http://www.internetworldstats.com

A Normalized Metric: Standardized Incidence Rate (SIR)



$$sir_{city} = \frac{ips_{city}}{n_{city}} \cdot 100{,}000$$

- **Used in the past to track cancer infection rate**
  - Above plot[1] shows the standardized incidence rate per state for lung and bronchus cancer across the United States in 2005
- **Our proposed metric is infection rate normalized for each 100,000 computers in each city**
  - Easy to understand whole numbers (1% is 1000)
  - Makes it possible to compare malicious activity rate across cities

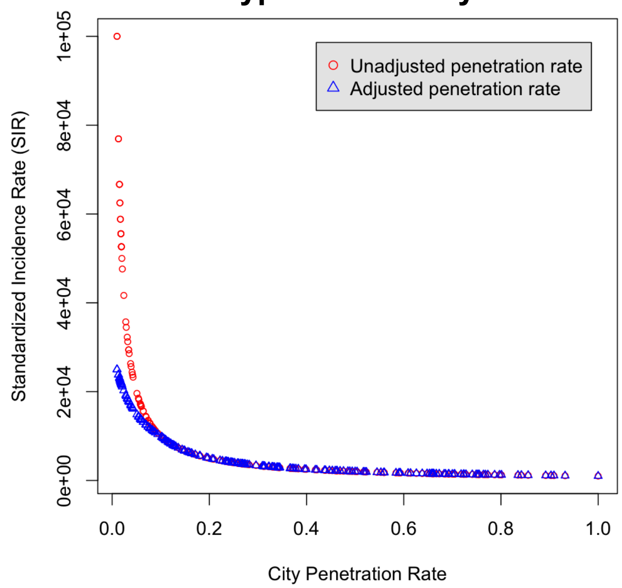
[1] Centers for Disease Control and Prevention. U.S. Cancer Statistics: An Interactive Atlas, March 2010. http://apps.nccd.cdc.gov/DCPC_INCA

Compensations for Data Flaws and Statistical Variability (1)

- **Uncertainties in Internet Penetration Rates**
  - SIR scores are highly sensitive for countries with low penetration rates
  - Higher measurement errors for countries with low rates
  - Developed countries have more steady rates than developing countries
  - Greater technological disparity between urban and rural areas in developing countries

Condition	Adjustment
$rate_{country} < 0.01$	Discard
$0.01 \le rate_{country} < 0.1$	Graduated amplification from 4x to 1x
$rate_{country} \ge 0.1$	Same

The Spread of SIR for a Hypothetical City



Compensations for Data Flaws and Statistical Variability (2)

- **Adding or removing one infected host (by chance) can dramatically change a city's SIR score under these conditions**
  - Small ips_{city}
  - Small n_{city}

$$sir_{city} = \frac{ips_{city}}{n_{city}} \cdot 100{,}000$$

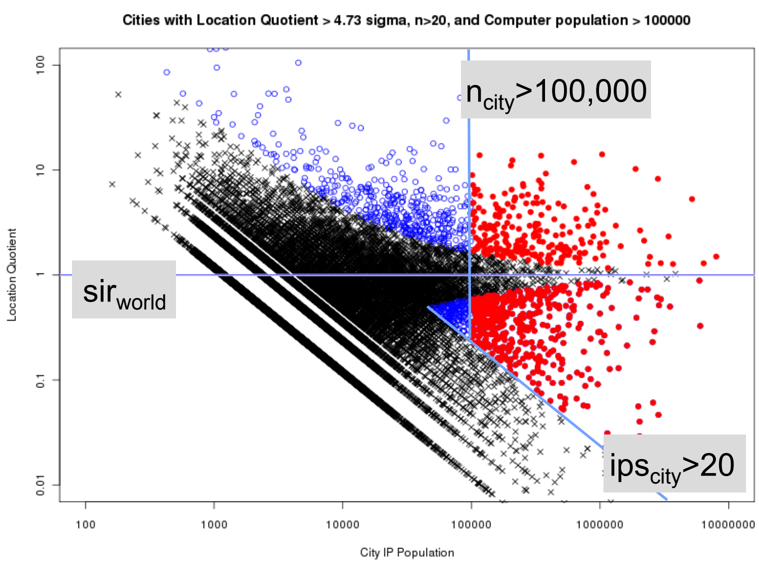
Example:

	Computer Population	Baseline	+1 Infection	Change in SIR
$City_A$	10,000	ips_{city} = 10 SIR = 100	ips_{city} = 11 SIR = 110	+10%
$City_B$	1,000,000	ips_{city} = 1000 SIR = 100	ips_{city} = 1001 SIR = 100.1	+.1%

- **To compensate for greater variability with smaller cities, EMBER only includes cities with at least 20 infections and 100,000 computers.**
  - ±10 infections should result in no more than ±5% change in SIR

Isn't This Just Hot or Cold Spot Analysis?

- **In cancer studies, the SIR is assumed to be binomially distributed around the global mean.**
- **Can city malicious activity be modeled similarly by assuming the probability of infection for any computer is the same?**

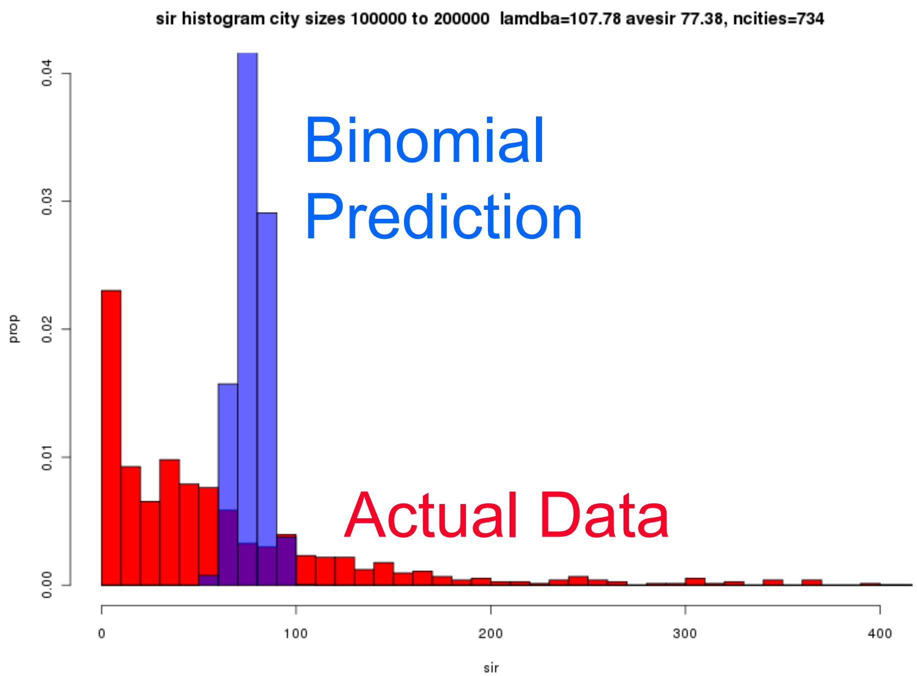


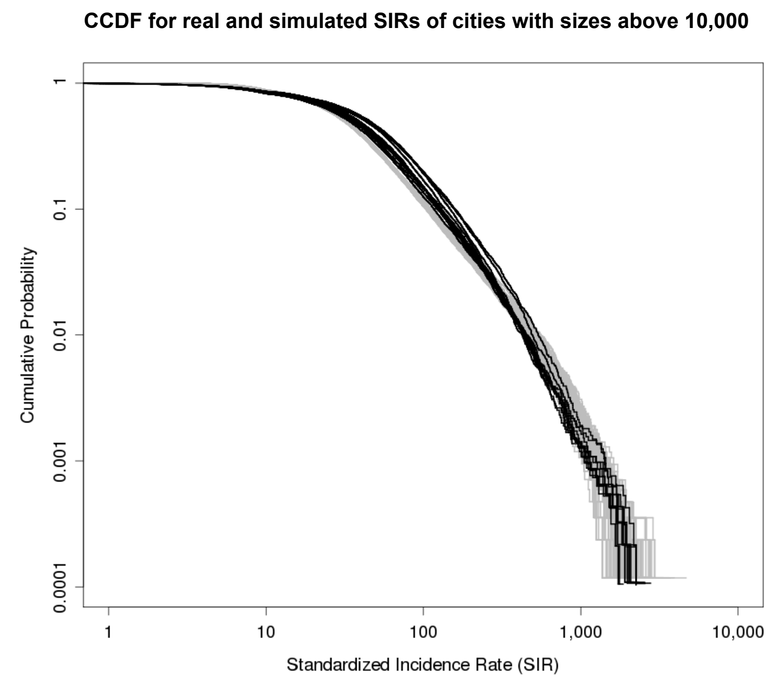
$$\mathrm{sir}_{\mathrm{city}} = \frac{\mathrm{ips}_{\mathrm{city}}}{\mathrm{n}_{\mathrm{city}}} \cdot 100{,}000$$

$$\sigma(\mathrm{sir}_{\mathrm{city}}) = \frac{\mathrm{sir}_{city}}{\sqrt{ips_{city}}}$$

Statistically significant cities with more or less malicious activity than expected if the distribution were binomial

We Discovered that SIRs are not Binomial but Have Long Tails





Experimental data shows that SIRs have a long-tail distribution, which is consistent with malware that spreads uniformly with a small probability (α) and spreads preferentially into cities proportional to the malicious activity already present with probability ($1- \alpha$).

SIR Ranking

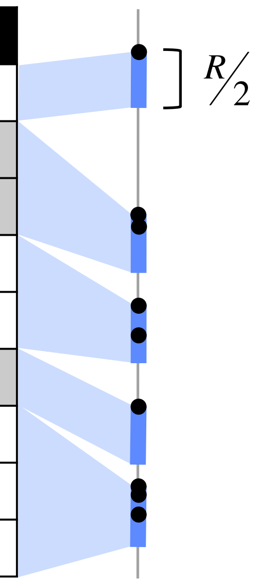
- **Goal: Assign identical ranks to cities with statistically equivalent SIR scores**
- **Compute cities' SIR confidence intervals (distribution-free) to determine the boundaries of equivalency**
  - Compute per-city 10-day interdecile range of SIR variability for all cities
  - Find the median 10-day interdecile range across cities (R)

Rank	City	SIR
1	Kaluga, RU	636.5820
2	Hyderabad, IN	534.2949
3	Lisbon, PT	533.6327
4	Sarajevo, BA	512.9266
5	Beijing, CN	508.8253
6	Vladimir, RU	484.3267
7	Vilnius, LT	466.8473
8	Taipei, TW	466.4215
9	Constanta, RO	463.8035

Simple Ranking

Rank	City	SIR
1	Kaluga, RU	636.5820
2	Hyderabad, IN	534.2949
2	Lisbon, PT	533.6327
3	Sarajevo, BA	512.9266
3	Beijing, CN	508.8253
4	Vladimir, RU	484.3267
5	Vilnius, LT	466.8473
5	Taipei, TW	466.4215
5	Constanta, RO	463.8035

EMBER Ranking

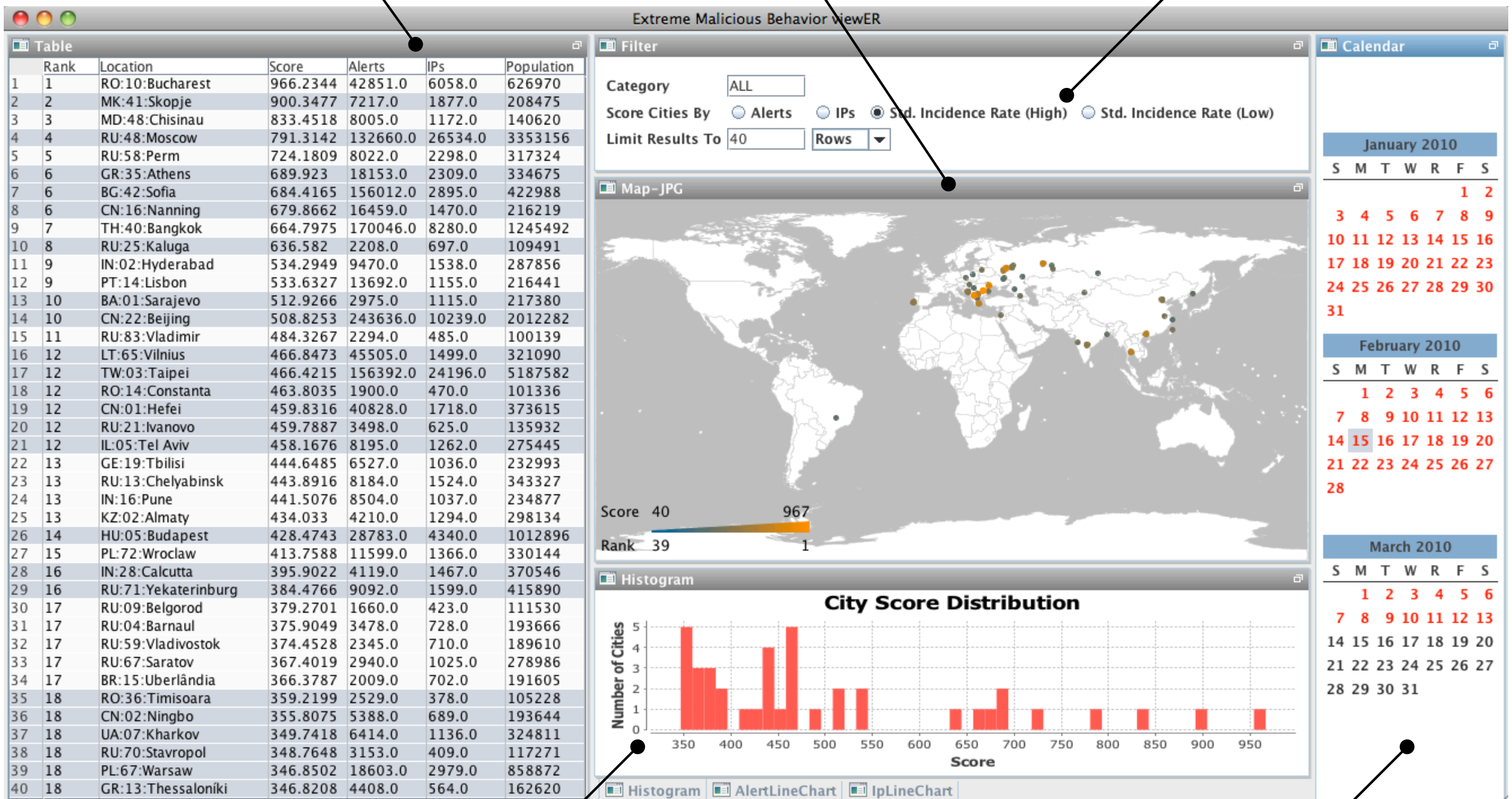


EMBER Display

top-ranked cities in detail

top-ranked cities in a world map

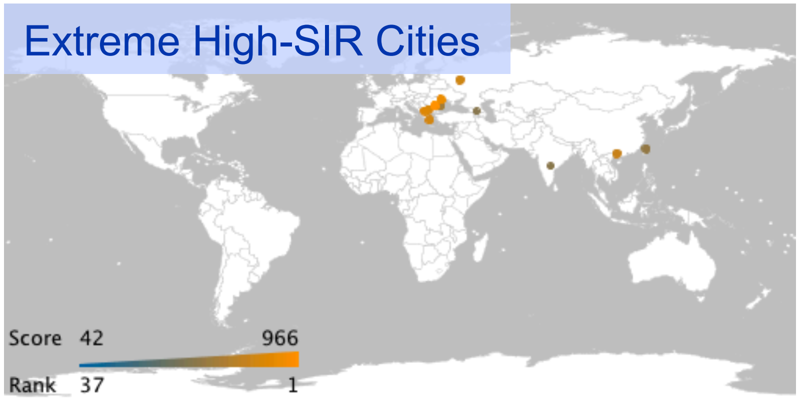
metric selection



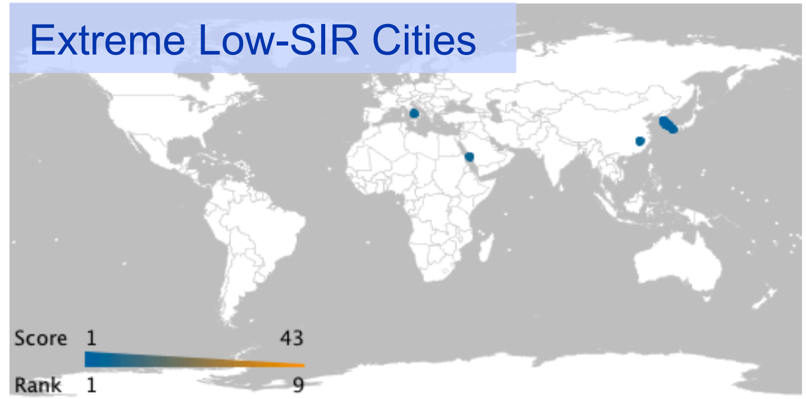
histogram of scores for cities shown

date selection

Useful Features of This World Map Display



	Rank	Location	Score	Alerts	IPs	Population
1	1	RO:10:Bucharest	965.9154	44209.0	6056.0	626970
2	2	MD:48:Chisinau	872.5643	5221.0	1227.0	140620
3	3	MK:41:Skopje	814.9658	8197.0	1699.0	208475
4	4	BG:42:Sofia	797.6585	152081.0	3374.0	422988
5	5	GR:35:Athens	780.7574	21325.0	2613.0	334675
6	5	RU:48:Moscow	775.9853	273984.0	26020.0	3353156
7	6	CN:16:Nanning	744.7479	13576.0	1736.0	233099
8	7	RO:14:Constanta	611.826	3334.0	620.0	101336
9	8	TW:03:Taipei	552.4346	145793.0	28658.0	5187582
10	9	IN:02:Hyderabad	517.9673	11132.0	1491.0	287856
11	9	GE:19:Tbilisi	513.7493	7545.0	1197.0	232993



	Rank	Location	Score	Alerts	IPs	Population
1	1	KR:12:Inchon	1.2306	510.0	25.0	2031444
2	1	KR:10:Busan	2.7782	1405.0	79.0	2843523
3	1	KR:19:Daejeon	2.8061	365.0	32.0	1140346
4	1	KR:15:Daegu	3.0242	700.0	60.0	1983935
5	1	JP:07:Kitakyushu	3.5849	163.0	27.0	753146
6	1	KR:12:Incheon	3.8888	1533.0	79.0	2031444
7	1	CN:03:Nanchang	4.0538	83.0	22.0	542692
8	1	KR:13:Bucheon	4.2578	170.0	28.0	657615
9	1	KR:21:Ulsan	4.2993	269.0	32.0	744295
10	1	IT:04:Napoli	4.6939	156.0	24.0	511299
11	1	SA:14:Jeddah	5.5954	303.0	43.0	768476

- **Highlight salient features in the dataset, not population centers**
- **Dot sizes and colors reveal regional variations**
- **Provide statistically valid ranking of per-city malicious activity**

Conclusions

- **We demonstrated an analytical approach toward developing a usable world map display of extreme malicious behavior**
  - Score cities by the Standardized Incidence Rate (SIR), which is the number of infections normalized by the local host population
  - Use publicly available data sources for estimating local host population
  - Apply careful adjustments to account for data flaws and statistical variability
  - Present a visualization that is as unbiased as possible

- **The high-SIR and low-SIR metrics are useful for exploring geographical variations**
  - Regions that are generally risky or well-protected
  - Regions that are targeted or avoided by specific threats

- **EMBER can be used on any IPv4 dataset. Higher-fidelity geo-location and population data could be integrated for better results.**