



NTT

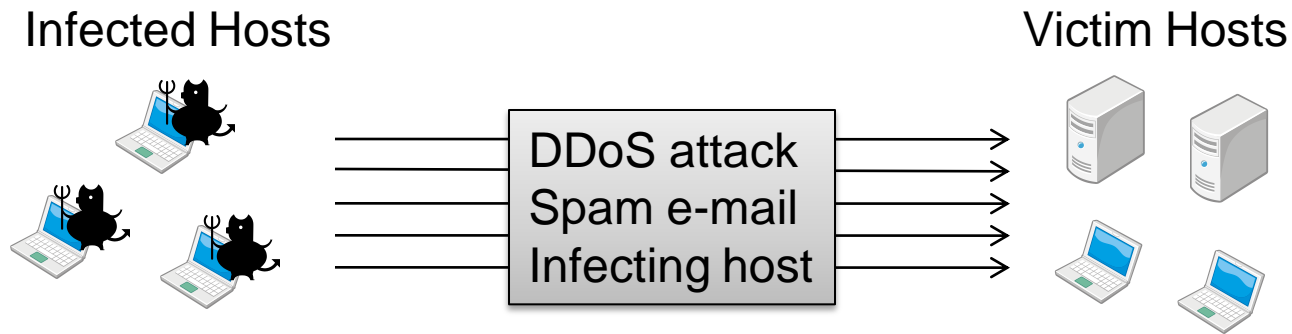
NTT Information Sharing Platform Laboratories

Extending Black Domain Name List by Using Co-occurrence Relation between DNS Queries

Kazumichi Sato, Keisuke Ishibashi,
Tsuyoshi Toyono, and Nobuhisa Miyake

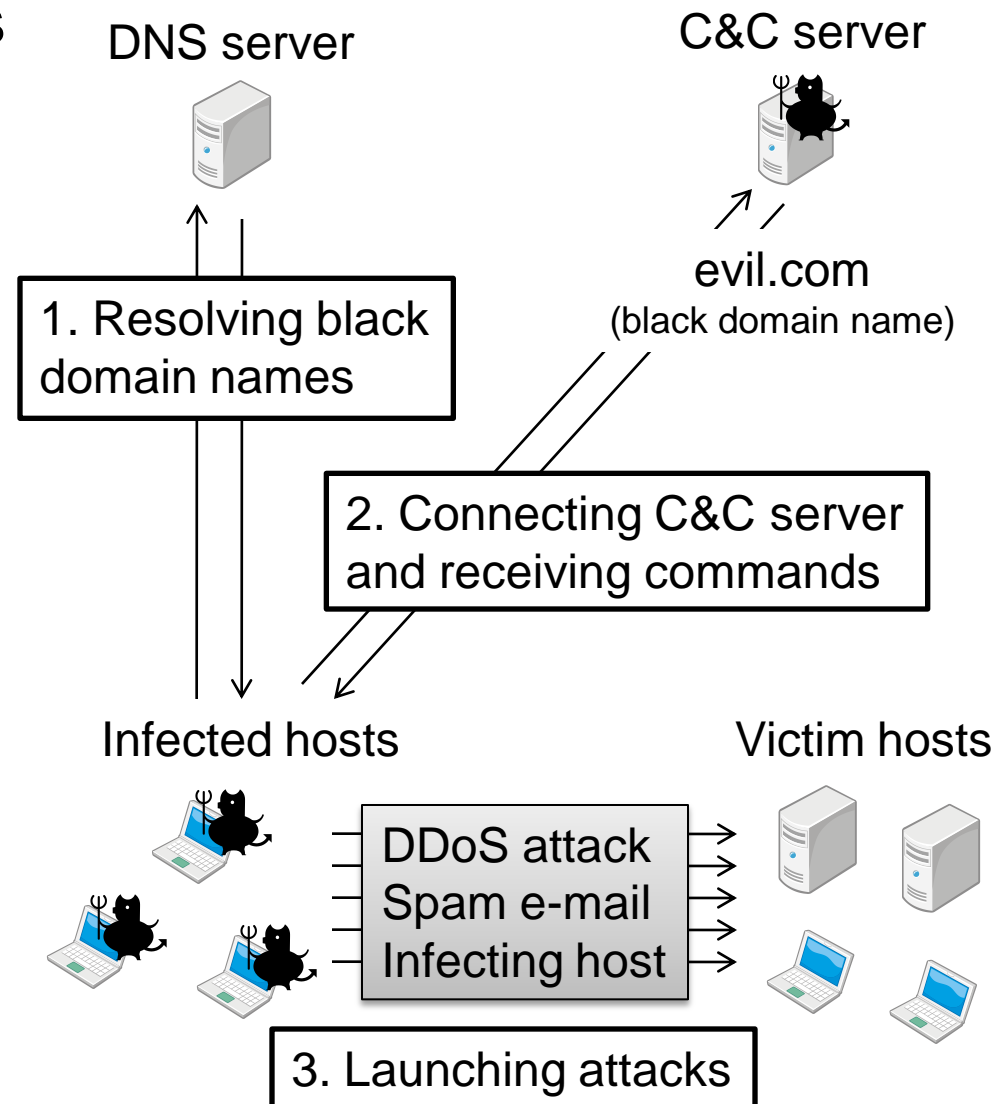
- Motivation
 - Method for detecting botnets by using blacklist
 - Coverage of blacklist
- Proposed Method
 - Extend blacklist by using co-occurrence relation
 - Problems of naively using co-occurrence relation
 - Eliminating popular domain names and heavy user effect
- Experimental Results
- Conclusion

- Botnet threats increasing
 - Launching DDoS attacks
 - Sending spam e-mail
 - Stealing personal information
 - Infecting other hosts
- Finding infected hosts and stopping malicious activities is necessary



Black Domain Name List

- Match black domain names with DNS queries to detect infected hosts
 - Bot sends DNS query to resolve domain name of C&C server
 - Black domain name list created by capturing and analyzing bots
- Block connections from infected hosts to C&C servers to stop malicious activities





- Blacklist does not cover all black domain names
 - Numerous new bots are observed every day, thus we cannot capture all bots
 - Some bots resolves many different black domain name , thus it is hard to maintain blacklist (e.g., Conficker worm)

- **Extending blacklist**

- Find unknown black domain names

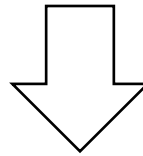
- Stop malicious activities by blocking connections from infected hosts to C&C servers

- Using extended blacklist, **find unknown infected hosts**

- Alert infected hosts to removing bot

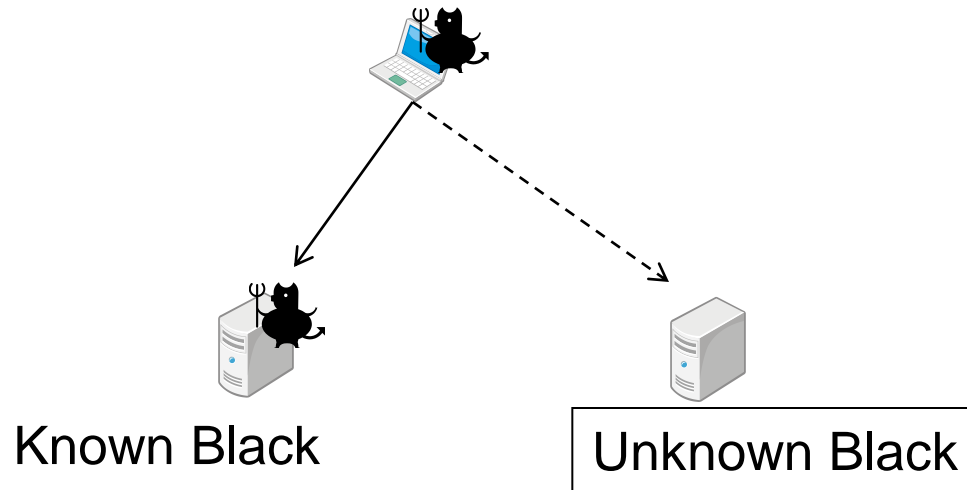
Assumption

- One bot resolves several black domain names
 - For redundancy of C&C servers

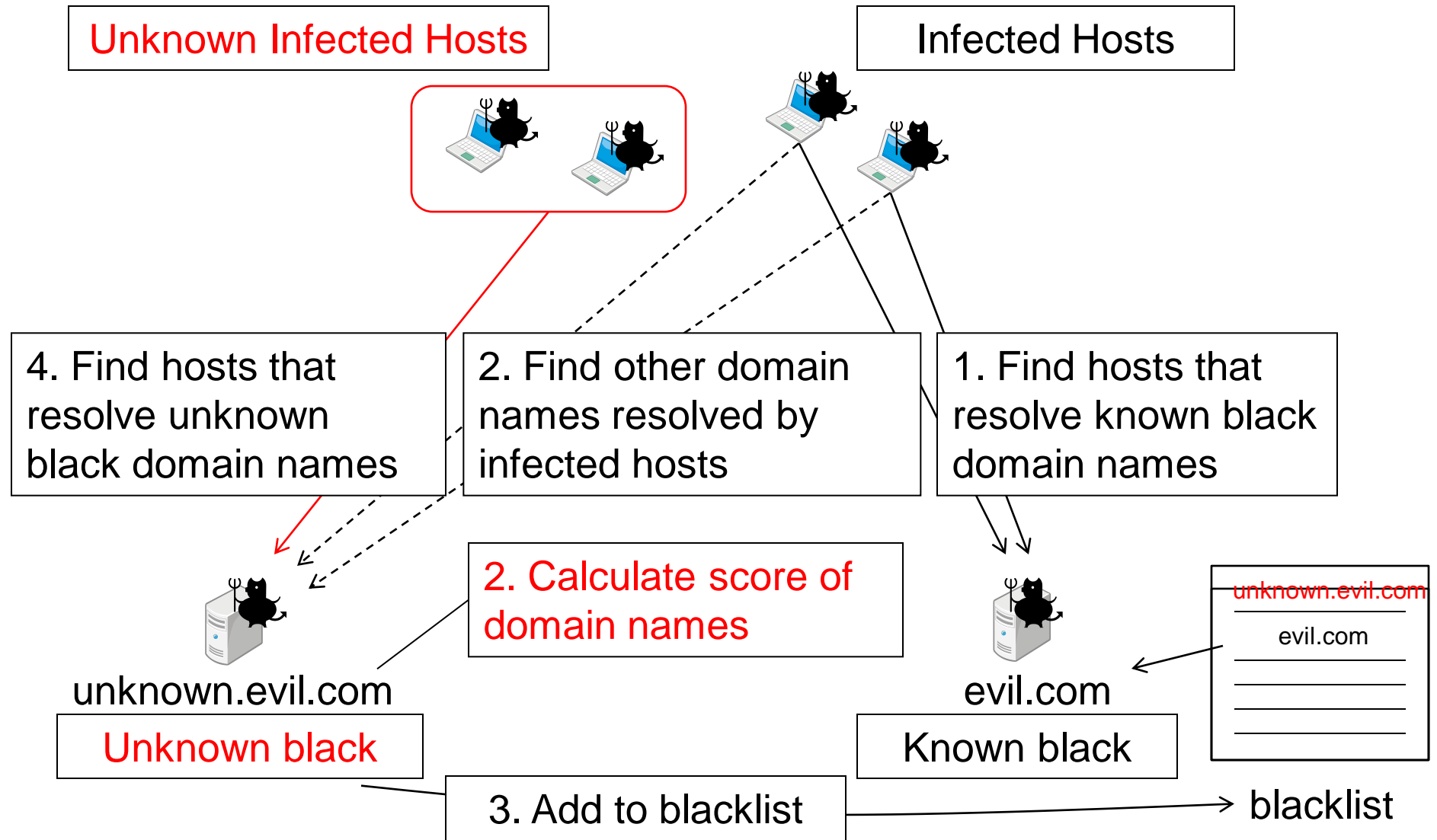


Assumption

If two domain names are resolved by the same host frequently and one is black, **the other is also black.**



Approach Overview



Naive Scoring Method

- Our assumption
 - If two domain names are resolved by the same host frequently and one is black, **the other is also black.**

Focus on **Co-occurrence relation**

- Scoring method by using co-occurrence relation

$$C(d_1, d_2) = \frac{\text{\# hosts that resolve } d_1 \text{ and } d_2}{\text{\# hosts that resolve } d_1 \text{ or } d_2}$$

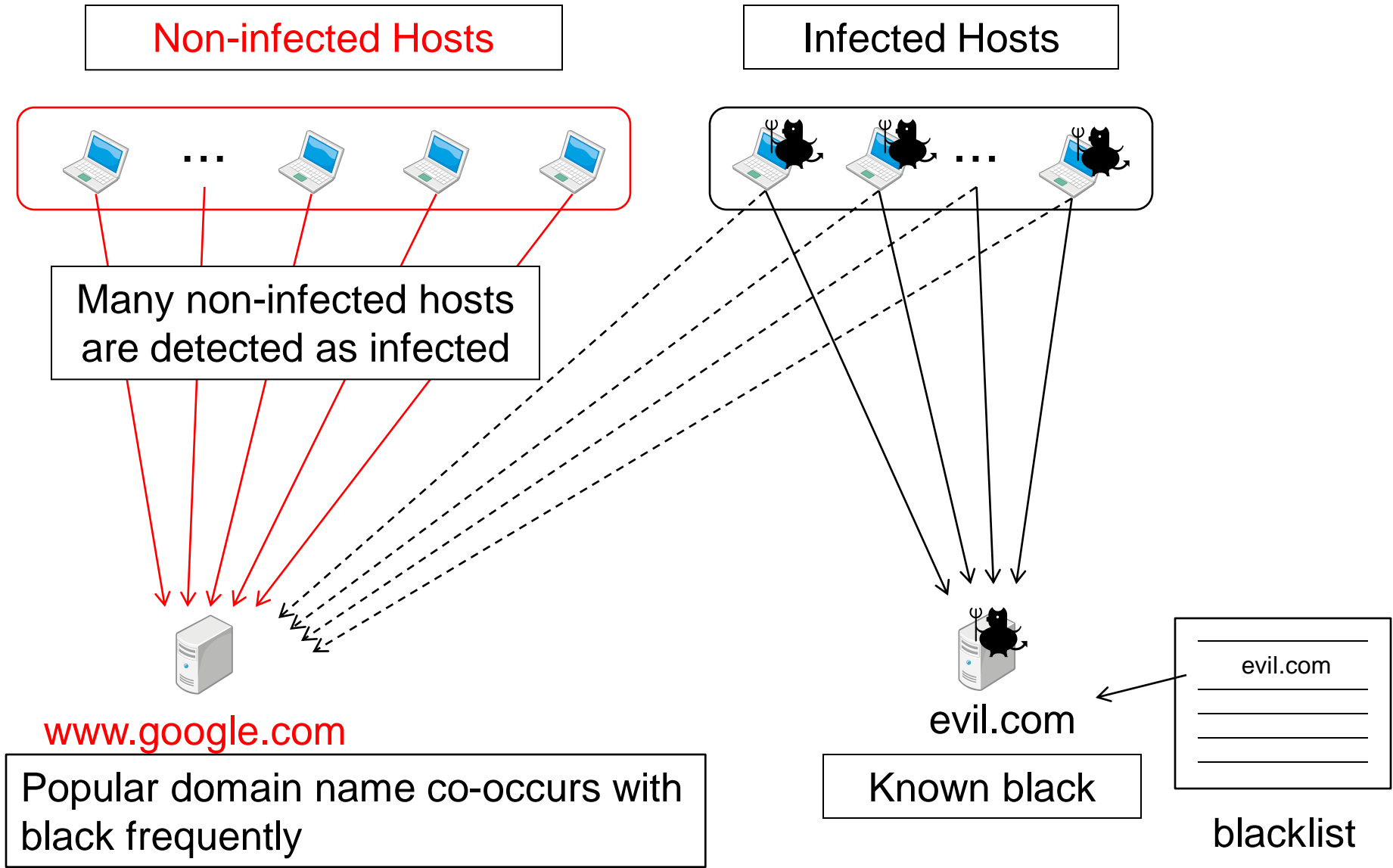
Co-occurrence rate

$$S(d) = \sum_{d_m \in \text{blacklist}} C(d_m, d)$$

Total co-occurrence rate with black domain names

If score is high, assume d is black

NTT Problem of Popular Domain Name



- Focus on number of non-infected hosts that resolve a domain name
 - Popular domain names are resolved by both infected and non-infected hosts
 - Black domain names are resolved by only infected hosts
- Define weight of number of non-infected hosts

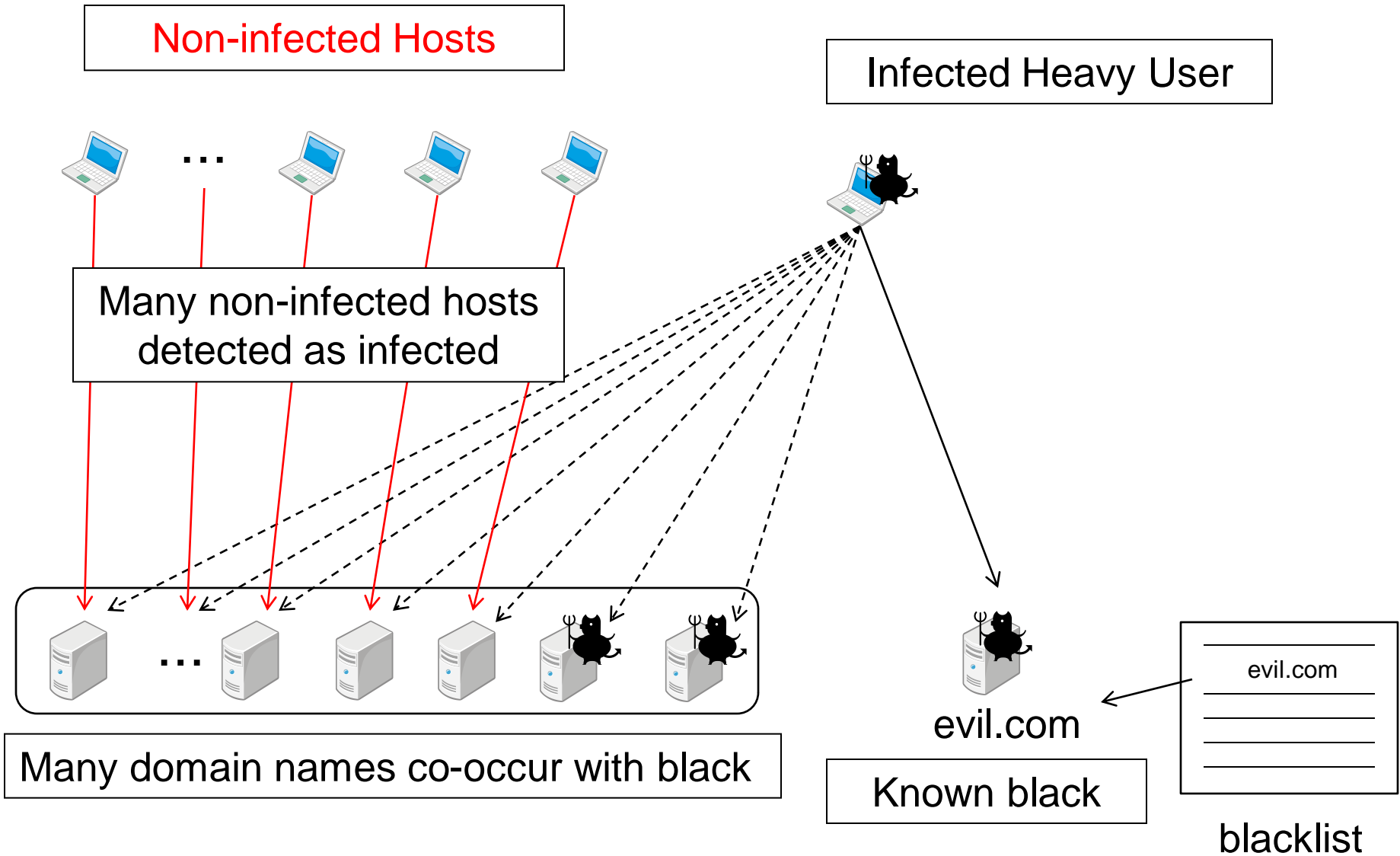
$$W(d) = \frac{\text{\# infected that resolve } d}{\text{\# infected that resolve } d + \text{\# non - infected that resolve } d}$$

If d is popular, $W(d)$ is relatively small

If d is popular, value is relatively large



Problem of Infected Heavy User



- Focus on number of domain names resolved by infected hosts
 - Add weight of number of queries to naive co-occurrence rate
- Weighted co-occurrence rate

$$C'(d_1, d_2) = \frac{\sum_{h \in \text{hosts that resolve } d_1 \text{ and } d_2} 1 / \# \text{ domain names resolved by } h}{\# \text{ hosts that resolve } d_1 \text{ or } d_2}$$

If host is heavy user, value is small and C' increases little

Proposed Scoring Method

Eliminate influence of infected heavy user

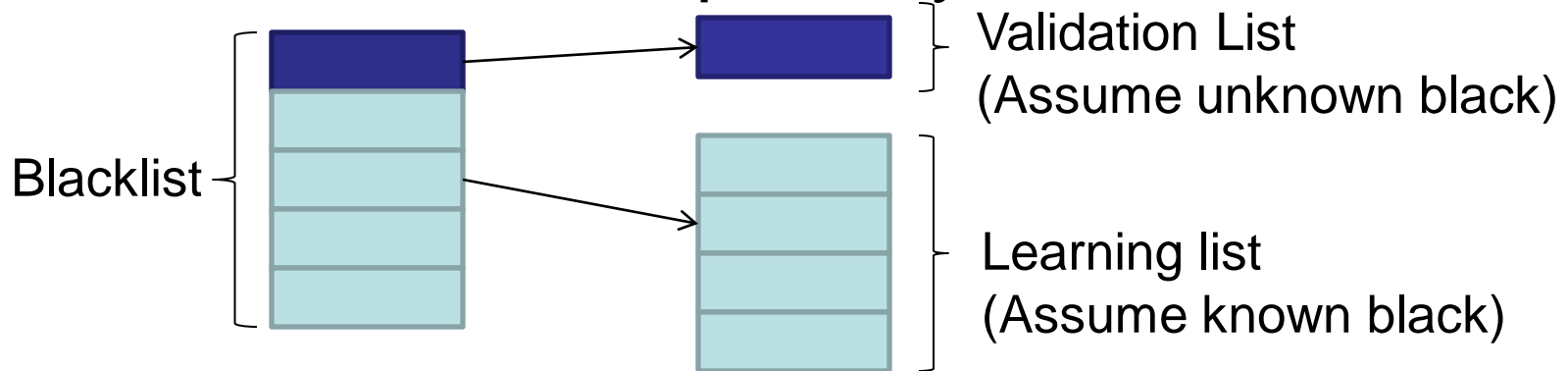
Even if d is resolved by infected heavy user, C' increases little

$$S'_w(d) = \left(\sum_{d_m \in \text{blacklist}} C'(d_m, d) \right) \times W(d)$$

Eliminate influence of popular domain name

If d is popular, $W(d)$ is small

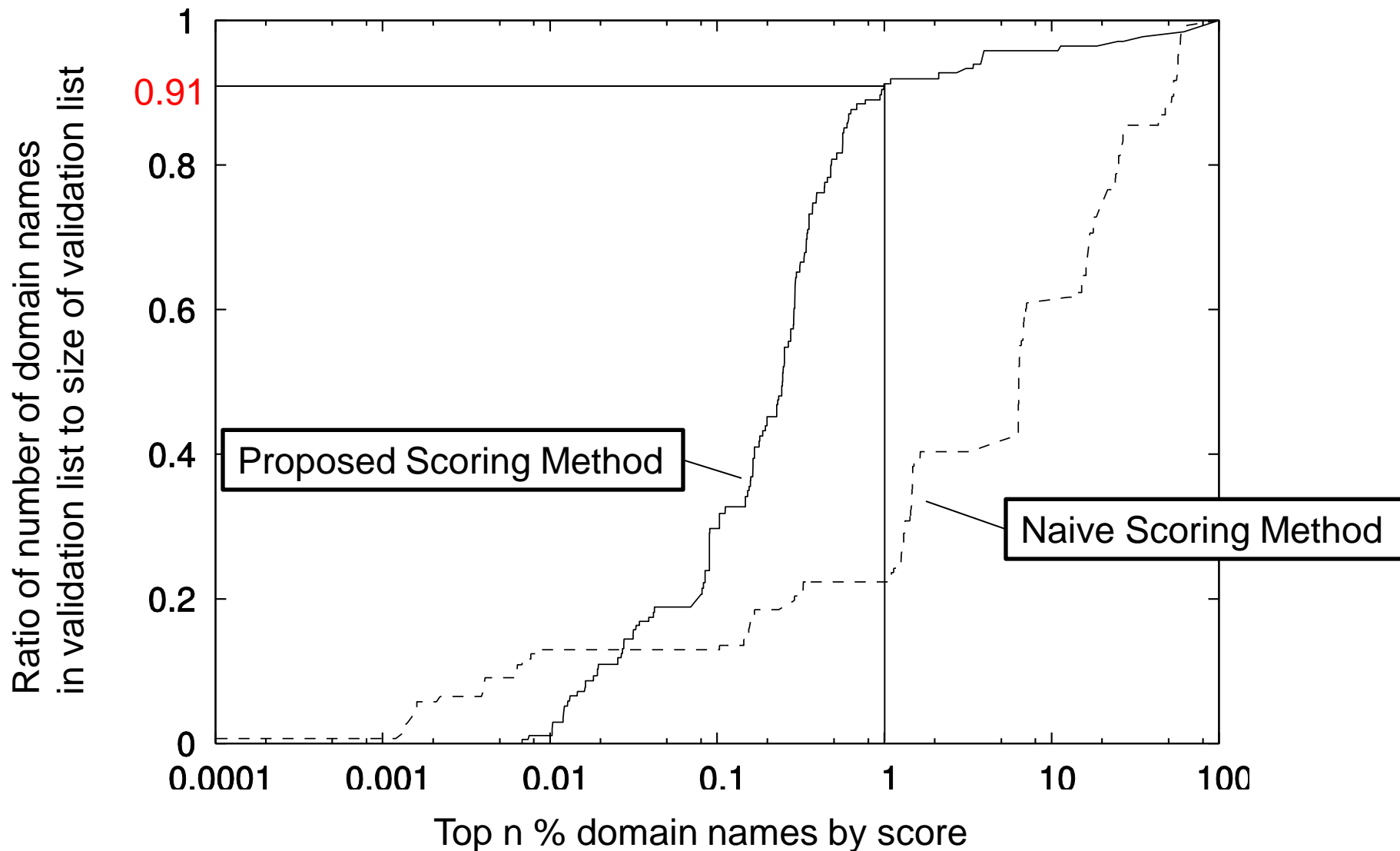
- Validated our assumption by cross validation



- Validated high-scored domain names
 - Applied proposed method to all domain names in known blacklist and classified 100 high-scored domain names as black, legitimate, or unclear
- Validated effectiveness of extended blacklist
 - Found hosts that resolved domain names in extended blacklist

- DNS traffic data
 - Captured during one hour in Feb. 2009
- Blacklist
 - Created by using honeypot during same period
 - Blacklist has about 270 domain names

Cross Validation Results



Classification Results

- Domain names for top 100 scores consisted of
 - 39% black, 4% legitimate, and 56% unclear
- Domain names for top 20 scores
 - 70% black
 - No legitimate domain names included

Score	Domain Name	Result
0.571	spy.nerashti.com	Black
0.571	bla.bihsecurity.com	Black
0.571	aaaaaaaaaaaaa.locop.net	Black
0.500	icq-msg.com	Black
0.319	mail.tiktikz.com	Black
0.300	x.zwned.com	Black
0.300	evolutiontmz.sytes.net	Unclear
0.300	dcom.anxau.com	Black
0.292	usa.lookin.at	Unclear
0.292	rewt.buyacaddi.com	Black

Score	Domain Name	Result
0.250	unkn0wn	Unclear
0.250	google-analitucs.com/loader/	Black
0.222	netspace.err0r.info	Unclear
0.203	win32.kernelupdate.info	Black
0.203	free.systemupdates.biz	Unclear
0.200	zjjdte.3322.org	Black
0.200	yklh.3322.org	Unclear
0.200	dr27.mcboo.com	Black
0.189	china.alwaysproxy.info	Black
0.167	home.najd.us	Black

- Some unclear domain names are suspicious
 - Domain name whose **subdomain differs from known black domain name**
 - ykln.3322.org (zjdtc.3322.org is known black)
 - Domain name with format “**<black>.<legitimate>**”
 - www.h7smcnrwlsgdn34fgv.info.<legitimate>
 - Domain name for **DNSBL lookups**
 - <IP address>.zen.spamhaus.org

- Rate of increase of number of unknown infected hosts is **only 3%**
 - Insufficient rate
 - Need to improve proposed method

- Proposed scoring method for finding unknown black domain names
- Found unknown black domain names and extended blacklist
 - Stop malicious activities by using extended blacklist more effectively
- Cannot find unknown infected hosts sufficiently
 - Improve method for finding unknown infected hosts as future work

Thank You