Finding New Malicious Domains Using Variational Bayes on Large-Scale Computer Network Data

Vojtěch Létal[2] Václav Šmídl[3] Tomáš Pevný[1,2] Petr Somol[1,3]
letal.vojtech@gmail.com smidl@utia.cas.cz tpevny@cisco.com psomol@cisco.com

[1]Cisco Systems, Cognitive Research
Karlovo náměstí 10, 12000 Prague, Czech Republic

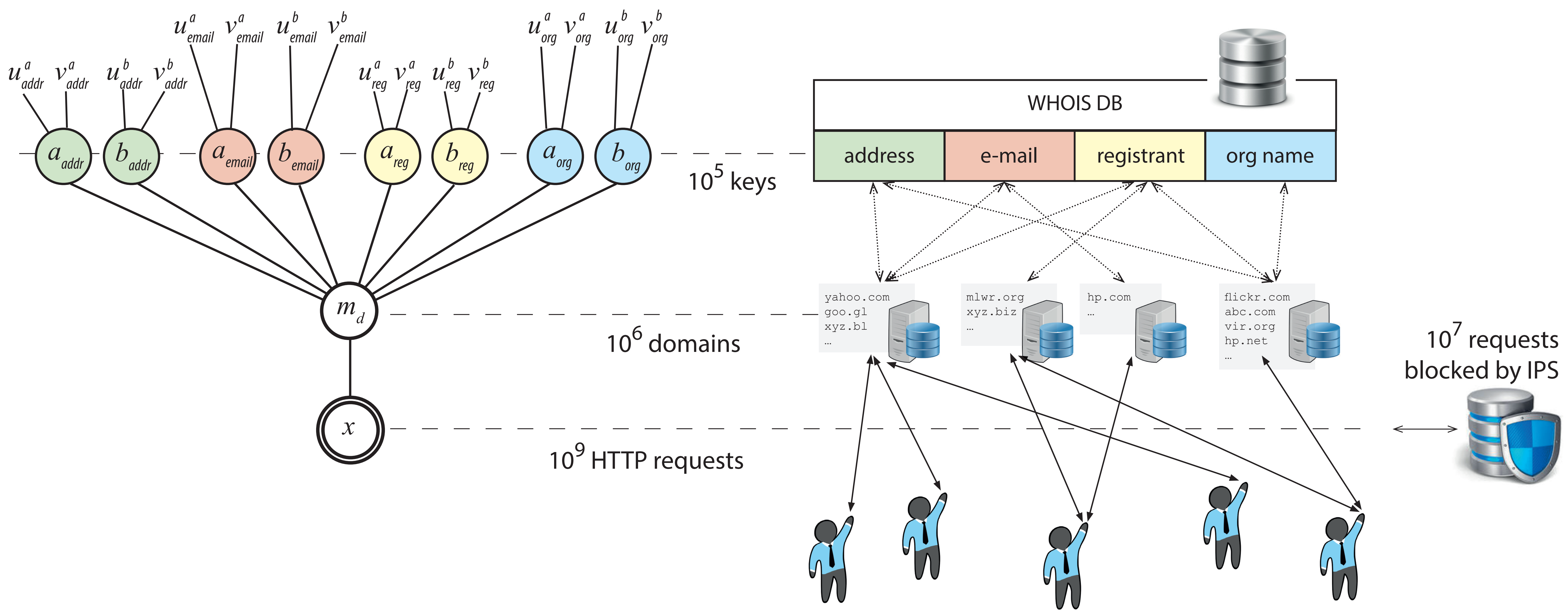
[2]Czech Technical University in Prague
FEL, Technická 2, 166 27 Prague 6, Czech Republic

[3]UTIA, Czech Academy of Sciences
Pod vodárenskou věží 4, Prague 8, Czech Republic

Abstract

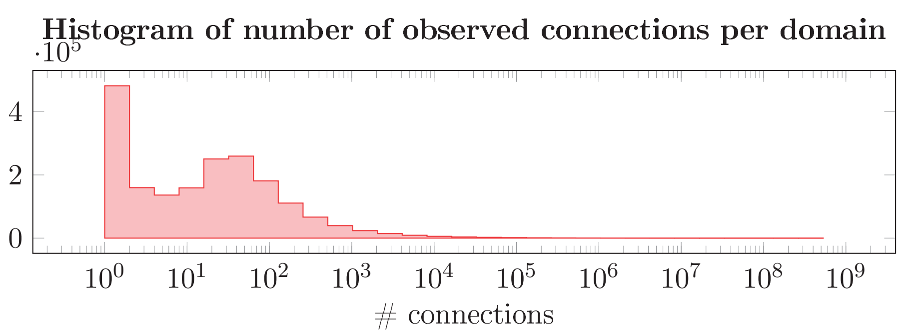
The common limitation in computer network security is the reactive nature of defenses. A new type of infection typically needs to be first observed live, before defensive measures can be taken. To improve the pro-active measures, we propose to utilize WHOIS database to model and estimate the probability of a domain name being used for malicious purposes from observed connections to other related domains. Model parameters are inferred by a Variational Bayes method. Its effectiveness is demonstrated on a large-scale network data with millions of domains and trillions of connections to them. The model enables *preventive blacklisting* in network security.

References:
[1] *Z. Ma, A. Leijon.* Bayesian estimation of beta mixture models with variational inference. *IEEE Trans. PAMI 33(11):2160-2173, 2011*
[2] *V. Šmídl, A. Quinn.* The variational Bayes method in signal processing. *Springer Science & Business Media, 2006.*



Problem

Modeling precise domain relations fails due to prevalently singular nature of observed connections:



Incomplete/garbled information about domain relations complicates things further.

Solution we found to work: model factorization.

Model

Let

$$p(x|m_d) = Bi(m_d)$$
$$p(m_d|a,b) = Beta(a_d, b_d)$$
$$a_d \approx a_{addr} \cdot a_{email} \cdot a_{reg} \cdot a_{org}$$
$$b_d \approx b_{addr} \cdot b_{email} \cdot b_{reg} \cdot b_{org}$$
$$p(a_*|u^a, v^a) = Gamma(u^a, v^a)$$
$$p(b_*|u^b, v^b) = Gamma(u^b, v^b)$$

Given training data $(d, b) \in T$ the complete model is:

$$p(M, A, B|T) \propto p(M, A, B, T) = p(T|M)p(M|A, B)p(A)p(B)$$

Inference

We approximate (assuming cond. indep.)

$$p(M, A, B|T) \approx q(M, A, B) = \prod_{d \in D} q(m_d) \prod_{l \in \mathcal{L}} q(a_l)q(b_l)$$

(where $\mathcal{L} = \mathcal{K}_{addr} \cup \mathcal{K}_{email} \cup \mathcal{K}_{reg} \cup \mathcal{K}_{org}$)

Minimize KL divergence by setting

$$\log q(b_l) \propto \mathbb{E}_{M,A,B\backslash b_l}[\log p(M, A, B|T)]$$
$$\log q(a_l) \propto \mathbb{E}_{M,A\backslash b_l,B}[\log p(M, A, B|T)]$$
$$\log q(m_d) \propto \mathbb{E}_{M\backslash m_d,A,B}[\log p(M, A, B|T)]$$

Using [1] recompute

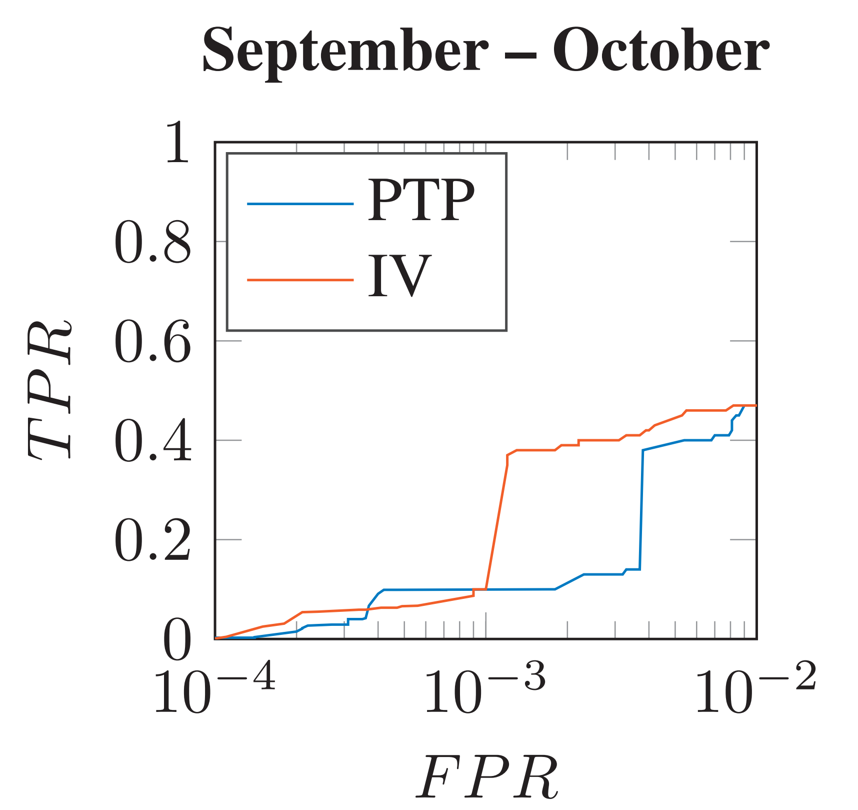
$$q(m_d) \sim \text{Beta}\left(\prod_{l \in k(d)} \widehat{a_l} + \sum_{x \in \mathcal{X}(d)} x,\ \prod_{l \in k(d)} \widehat{b_k} + \sum_{x \in \mathcal{X}(d)} (1-x)\right),$$

$$q(a_l) \sim \text{Gamma}\left(u_a + \sum_{\{d \in \mathcal{D}|l \in k(d)\}} \zeta_{d,k(d)},\ v_a - \sum_{\{d \in \mathcal{D}|l \in k(d)\}} \widehat{a_{k(d)\backslash l}} \widehat{\log m_d}\right)$$

$$q(b_l) \sim \text{Gamma}\left(u_b + \sum_{\{d \in \mathcal{D}|l \in k(d)\}} \zeta_{d,k(d)},\ v_b - \sum_{\{d \in \mathcal{D}|l \in k(d)\}} \widehat{b_{k(d)\backslash l}} \widehat{\log m_d}\right),$$

until convergence.

Experiment



NIPS 2015 Workshop on Advances in Approximate Bayesian Inference, Montreal, 11 Dec 2015