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

**Vojtěch Létal, Tomáš Pevný** *
Czech Technical University in Prague
Czech Republic
{letal.vojtech,pevnak}@gmail.com

**Václav Šmídl, Petr Somol***
UTIA, Czech Academy of Sciences
Czech Republic
{smidl,somol}@utia.cas.cz

## 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 have developed a method utilizing WHOIS database (database of entities that has registered a particular domain) to model relations between domains even those not yet used. The model estimates the probability of a domain name being used for malicious purposes from observed connections to other related domains. The parameters of the model is inferred by a Variational Bayes method, and its effectiveness is demonstrated on a large-scale network data with millions of domains and trillions of connections to them.

## 1 Motivation

In network security it is increasingly more difficult to react to influx of new threats. Reactive defense solutions based on constant anti-virus signature updates do not scale. We aim at revealing a significant fraction of zero-day (previously unknown) network threats pro-actively by inferring in advance the probability of domain names (domains) to be used for malicious purposes to which connections have not been yet observed. The idea is to use external information to link domains to those with already observed and investigated connections. The proposed model builds upon Bayesian statistics with unknowns inferred by variational methods. The model is applicable to estimate probability of being involved in malicious activities of any computer network entity, given a suitable external database is available to connect them. The concept is demonstrated on estimating fraction of malware-related connections to millions of domains related through records stored in publicly accessible WHOIS databases. The experimental results show the method is robust enough to be used in realistic large-scale computer network setting with significant amount of missing data.

### 1.1 Notation

Let $x \in 0, 1$ denote a realization of a random variable, which in this work denotes maliciousness of a network connection (e.g. request of a web browser) to a domain $d$ (e.g. `example.com`). A connection has value zero if it is benign and one if it is considered to be malicious by some intrusion detection system (IDS). The set of all observed connections used in the inference is denoted by $\mathcal{X}$, and the set of all domains $\mathcal{D}$. Since $x$ is a realization of a random variable, $\mathcal{X}$ contains many connections to the same domain $d$ denoted $\mathcal{X}(d)$. A mapping $d(x)$ returning a domain of a request $x$ will be useful below.

Let $m_d$ be a parameter of the Bernoulli distribution of domain $d$ determining the probability with which requests to $d$ are blocked. $\mathcal{M}$ is a set of $m_d$ of all domains, and those are the parameters we

*Tomáš Pevný and Petr Somol are also with Cisco Systems Inc.

want to infer. Individual domains $d$ are linked together by informations stored in WHOIS database, to which it is referred to as keys. $\mathcal{L}$ is the set of all keys in the WHOIS database. The role of keys in the model is symmetric, it is therefore assumed $\mathcal{L} = \mathcal{K}_o \cup \mathcal{K}_e \cup \mathcal{K}_n \cup \mathcal{K}_p$, where $\mathcal{K}_n, \mathcal{K}_o, \mathcal{K}_e, \mathcal{K}_p$ are sets of all registrants names, organizations, e-mail addresses and postal addresses. Note that if two keys from different categories share the same value, they are considered to be distinct, e.g. registrant with the name `Foo` and organization with the name `Foo` are considered to be two different keys in $\mathcal{L}$. The particular key (composed from values in the WHOIS database) of domain $d$ is a quadruple $k(d) = (k_n(d), k_o(d), k_e(d), k_p(d))$ of indexes from $\mathcal{L}$, corresponding to $k_n(d) \in \mathcal{K}_n$, $k_o(d) \in \mathcal{K}_o$, $k_e(d) \in \mathcal{K}_e$, and $k_p(d) \in \mathcal{K}_p$.

## 2 The model

The goal of the proposed system is to estimate probability, $m_d$, that a web request to domain $d$ (e.g. `example.com`) is related to a malicious activity (such requests are further called malicious). The system estimates these probabilities from frequencies with which individual requests are classified to be malicious by some Intrusion Prevention System (IPS) and uses informations about domains provided by the WHOIS database to link domains together. By virtue of this link, the system provides an estimate for domains to which few or none requests have been observed and / or inspected by the IDS. We assume that the WHOIS database provides for every domain a quadruple $k(d) = (k_{n(d)}, k_{o(d)}, k_{e(d)}, k_{p(d)})$, where $k_{n(d)}$ is the name of the registrant of domain $d$, $k_{o(d)}$ its organization, $k_{e(d)}$ its contact email address, and $k_{p(d)}$ the postal address. We model the above as follows. The probability of a request $x$ directed to domain $d$ being malicious follows Bernoulli distribution

$$p_b(x|m_d) = m_d^x(1-m_d)^{1-x}, \tag{1}$$

where it is assumed that all requests are independent. The parameter $m_d$, which is equal to the fraction of blocked request to $d$, is the parameter we would like to infer based on observed data $\mathcal{X}$ and the WHOIS record $\mathcal{K}$. To do so, we introduce a Beta prior on the value of $m_d$ in form

$$p_d = \frac{m_d^{a_{k(d)}-1}(1-m_d)^{b_{k(d)}-1}}{\beta(a_{k(d)}, b_{k(d)})}, \tag{2}$$

where parameters $a_{k(d)}$ and $b_{k(d)}$ are parameters of the Beta-prior depending on the key $k(d)$, which is an unique quadruple retrieved from the WHOIS database. Formulation (2) can be possibly used to infer all $m_d \in \mathcal{M}$ using maximum likelihood approach, but it makes Beta priors $a_{k(d)}$, $b_{k(d)}$ unique to each quadruple $k(d)$, which means that two records differing in one or more sub-keys (e.g. postal address) have to be treated independently. Moreover both priors would need to be guessed sufficiently well, which would be very difficult. To resolve both problems, priors $a_{k(d)}$ and $b_{k(d)}$ are factorised such that each factor depends only on the portion of the key (e.g. e-mail address) and a Gamma prior is put on each factor [4]. Specifically,

$$a_{k(d)} = a_{k_n(d)} a_{k_o(d)} a_{k_e(d)} a_{k_p(d)} \text{ and } b_{k(d)} = b_{k_n(d)} b_{k_o(d)} b_{k_e(d)} b_{k_p(d)}, \tag{3}$$

where $k_n(d) \in \mathcal{K}_n$, $k_o(d) \in \mathcal{K}_o$, $k_e(d) \in \mathcal{K}_e$, $k_p(d) \in \mathcal{K}_p$ with $\mathcal{K}_n$, $\mathcal{K}_o$, $\mathcal{K}_e$, and $\mathcal{K}_p$ being set of all unique names, organizations, e-mail addresses, and postal addresses respectively. To simplify notation, we introduce $\mathcal{L} = \mathcal{K}_o \cup \mathcal{K}_e \cup \mathcal{K}_p \cup \mathcal{K}_n$ (if a same value of key appears in for example $\mathcal{K}_o$ and $\mathcal{K}_e$, they are treated as two different keys). With this notation we index $a_{k_n(d)}, a_{k_o(d)}, a_{k_e(d)}$, and $a_{k_p(d)}$ as $a_l$, $l \in \mathcal{L}$, and we can write $a_{k(d)} = \prod_{l \in k(d)} a_l$ and $b_{k(d)} = \prod_{l \in k(d)} b_l$. Finally, we introduce Gamma prior on $a_l$ and $b_l$, $l \in \mathcal{L}$ yielding to he final model

$$p(a_l) = \frac{{v_a}^{u_a}}{\Gamma(u_a)} a_l^{u_a-1} e^{-v_a a_l} \text{ and } p(b_l) = \frac{{v_b}^{u_b}}{\Gamma(u_b)} b_l^{u_b-1} e^{-v_b b_l}. \tag{4}$$

The final joint-probability function of the model is

$$\begin{aligned} p(\mathcal{M}, \mathcal{A}, \mathcal{B}|\mathcal{X}, \theta) = &\prod_{x \in \mathcal{X}} m_{d(x)}^x (1-m_{d(x)})^{1-x} \prod_{d \in \mathcal{D}} \frac{m_d^{a_{k(d)}-1}(1-m_d)^{b_{k(d)}-1}}{\beta\left(a_{k(d)}, b_{k(d)}\right)} \cdot \\ &\prod_{l \in \mathcal{L}} \frac{{v_a}^{u_a}}{\Gamma(u_a)} a_l^{u_a-1} e^{-v_a a_l} \frac{{v_b}^{u_b}}{\Gamma(u_b)} b_l^{u_b-1} e^{-v_b b_l}, \end{aligned} \tag{5}$$

where $a_{k(d)}$ and $b_{k(d)}$ are from (3).

**Algorithm 1** Variational Bayes inference of model parameters $\mathcal{M} = \cup_{d\in\mathcal{D}} m_d$, $\mathcal{A} = \cup_{l\in\mathcal{L}} a_l$, $\mathcal{B} = \cup_{l\in\mathcal{L}} b_l$. The convergence criterion is the change of expected value of $m_d$.

1. choose the parameters $\Theta = (u_a, v_a, u_b, v_b)$
2. choose initial estimates of $q(a_l)$ and $q(b_l)$
3. **For** it = 1 to max_iterations
   (a) $\forall d \in \mathcal{D}$ recompute $q(m_d)$
   $\forall d \in \mathcal{D}$ evaluate $\widehat{\log m_d}$, $\widehat{\log(1-m_d)}$
   (b) $\forall a_l \in \mathcal{A}$ recompute $q(a_l)$
   $\forall a_l \in \mathcal{A}$ evaluate $\widehat{a_l}$, $\widehat{\log a_l}$
   (c) $\forall b_l \in \mathcal{B}$ recompute $q(b_l)$
   $\forall b_l \in \mathcal{B}$ evaluate $\widehat{b_l}$, $\widehat{\log b_l}$
   (d) **If** convergence_criteria $<$ threshold **Then** break

## 3 Inference of parameters by Variational Bayes

Since an analytical solution of (5) does not exist due to the integral of the beta distribution not having a closed-form solution, a Variational Bayes method [5, 1] approximating posterior distribution implied by (5) using a product of marginal distributions is evaluated

$$p(\mathcal{M}, \mathcal{A}, \mathcal{B}|\mathcal{X}, \theta) \approx q(\mathcal{M}, \mathcal{A}, \mathcal{B}) = \prod_{d\in\mathcal{D}} q(m_d) \prod_{l\in\mathcal{L}} q(a_l)q(b_l). \tag{6}$$

The Variational Bayes method minimizes the KL-divergence between the approximation (6) and the joint pdf (5) by iteratively optimizing individual marginals $q(m_d)$, $q(a_l)$, and $q(b_l)$ while keeping all other marginals fixed. After sufficient number of iterations the algorithm converges to a local minimum. The algorithm is outlined in Algorithm 1 with the marginals[1] being

$$q(m_d) \sim \text{Beta}\left(\prod_{l\in k(d)} \widehat{a_l} + \sum_{x\in\mathcal{X}(d)} x,\ \prod_{l\in k(d)} \widehat{b_k} + \sum_{x\in\mathcal{X}(d)} (1-x)\right), \tag{7}$$

$$q(a_l) \sim \text{Gamma}\left(u_a + \sum_{\{d\in\mathcal{D}|l\in k(d)\}} \zeta_{d,k(d)},\ v_a - \sum_{\{d\in\mathcal{D}|l\in k(d)\}} \widehat{a_{k(d)\setminus l}}\widehat{\log m_d}\right) \tag{8}$$

$$q(b_l) \sim \text{Gamma}\left(u_b + \sum_{\{d\in\mathcal{D}|l\in k(d)\}} \zeta_{d,k(d)},\ v_b - \sum_{\{d\in\mathcal{D}|l\in k(d)\}} \widehat{b_{k(d)\setminus l}}\widehat{\log m_d}\right), \tag{9}$$

$$\tag{10}$$

where

$$\zeta_{d,k(d)} = \left[\psi(\widehat{a_{k(d)}} + \widehat{b_{k(d)}}) - \psi(\widehat{a_{k(d)}}) + \widehat{b_{k(d)}}\psi'(\widehat{a_{k(d)}} + \widehat{b_{k(d)}})(\widehat{\log b_{k(d)}} - \log\widehat{b_{k(d)}})\right], \tag{11}$$

$a_{k(d)\setminus l} = \prod_{l'\in k(d)\wedge l'\neq l} a_{l'}$, $b_{k(d)\setminus l} = \prod_{l'\in k(d)\wedge l'\neq l} b_{l'}$, and the $\hat{}$ denotes expected value of the variable. Exact variational marginals do not have an analytical solutions due to the intractability of the Beta distribution. Therefore we have used approximation proposed in [4] to obtain $q(a_l)$ and $q(b_l)$. The approximations are based on Taylor expansion of the logarithm of the Beta distribution in $\widehat{a_{k(d)}}$, $\widehat{b_{k(d)}}$ from previous iteration. Since the Taylor expansion is valid only for $\widehat{a_{k(d)}}$, $\widehat{b_{k(d)}} > 1$, both variables are set to one if they are smaller.

## 4 Experimental results

The proposed method was evaluated on data obtained from web-logs Cisco's Cloud Web Security [3], which is a cloud web proxy scrutinizing the traffic for the presence of known malware and

[1] Derivation of the marginals can be found in the supplemental material.

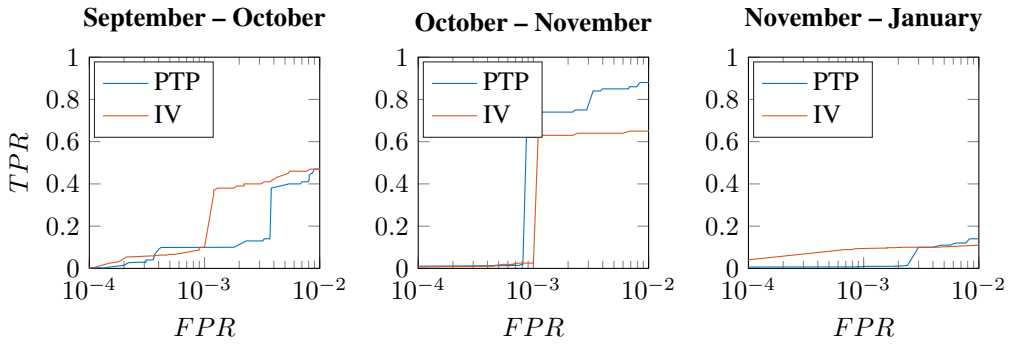


Figure 1: Comparison of the proposed method (IV) with the Probabilistic Threat Propagation (PTP).

other threats. For the experiments we have used traffic from the first week of September, October, November 2014, and January and February 2015. Each week of traffic contained approximately $7.5 \cdot 10^9$ connections (size of the set $\mathcal{X}$) out of which approximately $20 \cdot 10^3$ were deemed as malicious) to approximately $2 \cdot 10^6$ domains (size of the set $\mathcal{D}$) with approximately $10^5$ keys retrieved from the WHOIS database.

The data allows a natural division into training and testing data, where we have used the data from a previous month for training (inferring values in sets $\mathcal{A} = \cup_{l \in \mathcal{L}} a_l$, $\mathcal{B} = \cup_{l \in \mathcal{L}} b_l$, and $\mathcal{M} = \cup_{d \in \mathcal{D}} m_d$) and the data from the next month for the evaluation.

The proposed method was compared to Probabilistic Threat Propagation [2], which is a method that propagates a probability of certain network node being malicious based on connection graph. The method extrapolates maliciousness of "tips", which are domains used for malicious purposes, to other connected domains, which were domains sharing at least one key in WHOIS database. In our experiments "tips" were domains with at least 20% blocked connection, which was a fraction determined to avoid trivial false positives like `yahoo.com`.

Figure 1 shows ROC curves when the inference was done on data captured in September, October, and November, and evaluated on October, November, and January respectively. The ROC curve was obtained by changing the threshold on $m_d$ from which the domain would be considered malicious and counting correctly classified flows (blocked vs. not blocked). The false positive rate is drawn in logarithmic scale, because in security applications only very low false positive rate are interesting. Therefore only the region from zero to one percent false positive rate is shown to decide which algorithm is better. We observe that ROC curves of both methods intersect, but that of the proposed method is generally above (better) in the region of interest. Moreover, the proposed method does not require known "tips", which is an important feature for practical deployment, as it can be executed autonomously.

## 5 Conclusion

We have defined and verified a Bayesian model to infer probability of a network entity being involved in malicious activities. The important feature for practice is that the model propagates probability from entities with observed connections to those without using external information relating entities together. The model was instantiated to enable preventive blacklisting of yet unobserved domains using information about observed HTTP request blocks and domain registration records in the WHOIS database. The scalability of the model was shown on modeling millions of domains using trillions of web requests.

## 6 Acknowledgement

The work of V. Létal, P. Somol and T. Pevný has been partially supported by Czech Science Foundation project 15-08916S.

## References


[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] Kevin M Carter, Nwokedi Idika, and William W Streilein. Probabilistic threat propagation for malicious activity detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 2940–2944. IEEE, 2013.

[3] Inc. Cisco Systems˙ Cloud web security, 2015. ¡http://www.cisco.com/c/en/us/products/security/cloud-web-security/index.html¿.

[4] Zhanyu Ma and Arne Leijon. Bayesian estimation of beta mixture models with variational inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2160–2173, 2011.

[5] Václav Šmídl and Anthony Quinn. *The variational Bayes method in signal processing*. Springer Science & Business Media, 2006.

# Suplemental material for Finding New Malicious Domains Using Variational Bayes on Large-Scale Computer Network Data

**Vojtěch Létal, Tomáš Pevný** *
Department of Computer Science
Czech Technical University in Prague
Czech Republic
{letal.vojtech,pevnak}@gmail.com

**Václav Šmídl, Petr Somol***
Institute of Information Theory and Automation
Czech Academy of Sciences
Czech Republic
{smidl,somol}@utia.cas.cz

## 1 Recapitulation of variational Bayes to introduce notation

The recapitulation of Variational Bayes follows [1] page 464. To simplify the notation, we assume $p(z, d)$ to be the joint distribution of data $d$ and parameters $z$ we are interesting in. Furthermore, $p(x)$ is the marginal distribution of the data and $q(z)$ is a distribution by which we want to approximate $p(z|x)$ (distribution of parameters given data). Naturally it is assumed $z$ to be multi-dimensional with $k$ components. It holds that

$$\begin{aligned}\log p(x) &= \int_{\mathcal{Z}} q(z) \log p(x) dz \\ &\underset{(1)}{=} \int_{\mathcal{Z}} q(z) \log \left( \frac{p(x,z)}{p(z|x)} \frac{q(z)}{q(z)} \right) dz \\ &= \int_{\mathcal{Z}} q(z) \log \frac{p(x,z)}{q(z)} dz - \int_{\mathcal{Z}} q(z) \log \frac{p(z|x)}{q(z)} dz \\ &\underset{(2)}{=} \mathcal{E}(q(z)) + \mathrm{D}_{\mathrm{KL}} \left( q(z) || p(z|x) \right),\end{aligned} \tag{1}$$

where in $(1)$ we have used the fact that $p(x, z) = p(z|x)p(x)$ and in $(2)$ we have recognized KL-divergence in the second term and denoted the first one by $\mathcal{E}(q(z)) = \int_{\mathcal{Z}} q(z) \log \frac{p(x,z)}{q(z)} dz$.

The KL-divergence $\mathrm{D}_{\mathrm{KL}} \left( q(z) || p(z|x) \right)$ measures the closeness of $q(z)$ and $p(z|x)$. While it is not a true distance, it is non-negative and equal to zero iff $q(z) = p(z|x)$. Since the KL-divergence is always positive, by maximizing $\mathcal{E}(q(z))$ we minimize the $\mathrm{D}_{\mathrm{KL}} \left( q(z) || p(x|z) \right)$ for which we usually do not have a closed form solution. Thus finding the best approximation $q(z)$ amounts to maximize $\mathcal{E}(q(z))$.

*Tomáš Pevný and Petr Somol are also with Cisco Systems Inc.

Factorized models assume that $q(z)$ is a product of independent probability distributions, i.e. $q(z) = \prod_i q_i(z_i)$. With this assumption one can readily write

$$\begin{aligned}\mathcal{E}(q(z)) &= \int_{\mathcal{Z}} q(z) \log \frac{p(x,z)}{q(z)} dz \\ &= \int_{\mathcal{Z}_j} q_j(z_j) \left( \int_{\mathcal{Z}\setminus j} \prod_{i \neq j} q_i(z_i) \log p(x,z) dz_{\setminus j} \right) dz_j - \sum_i \int_{\mathcal{Z}} q(z) \log q_i(z_i) dz,\end{aligned} \tag{2}$$

where the notation $\setminus j$ means with respect or over to all variables except $j$. Let now introduce

$$\log \tilde{p}_j(x,z) = \int_{\mathcal{Z}\setminus j} \prod_{i \neq j} q_i(z_i) \log p(x,z) dz_{\setminus j} = \mathrm{E}_{\mathcal{Z}\setminus j \sim q_{\setminus j}} \left[ \log p(x,z) \right] \tag{3}$$

and denote $\eta_i = \int_{\mathcal{Z}_i} q_i(z_i) \log q_i(z_i) dz_i$. The the above can be simplified to

$$\begin{aligned}\mathcal{E}(q(z)) &= \int_{\mathcal{Z}_j} q_j(z_j) \log \frac{\tilde{p}_j(x,z)}{q_j(z_j)} dz_j - \sum_{i \neq j} \eta_j \\ &= -\mathrm{D}_{\mathrm{KL}} \left( q_j(z_j) || \tilde{p}_j(x,z) \right) - \sum_{i \neq j} \eta_j\end{aligned} \tag{4}$$

Now imagine that we optimize $\mathcal{E}(q(z))$ only with respect to $q_j(z_j)$. The the sum on the right-hand is independent and the KL term is maximized if the KL-divergence is equal to zero, which is when $q_j(z_j) = \tilde{p}_j(x,z)$.

## 2 Extended factorized approximation method

The problem we face in our work is that the expectation $\mathrm{E}_{\mathcal{Z}\setminus j \sim q_{\setminus j}} \left[ \log p(x,z) \right]$ is not tractable due to the integration of the logarithm of the beta function. The extended factorized approximation proposes to find a lower bound on $\mathcal{E}(q(z))$ and maximise this lower bound instead of $\mathcal{E}(q(z))$. By maximising the lower bound iteratively we can still asymptotically reach the optimum. Specifically, Ref. [2] search a lower bound on $p(x,z)$ (further denoted as $\tilde{p}(x,z)$) such that it holds that

$$\int q(z) \log p(x,z) dz \geq \int q(z) \log \tilde{p}(x,z) . dz \tag{5}$$

The last inequality trivially holds iff $p(x,z) \geq \tilde{p}(x,z)$.

The extended factorized approximation method maximizes $\int_{\mathcal{Z}} q(z) \log \frac{\tilde{p}(x,z)}{q(z)} dz$ instead of $\mathcal{E}(q)$ by using the variational Bayes method. In every iteration, the lower bound (or envelope) is updated, by which the same solution can be reached as if the original $\mathcal{E}(q)$ would be optimised.

#### Derivation of marginals

The joint pdf function is

$$\begin{aligned}p(\mathcal{M}, \mathcal{A}, \mathcal{B} | \mathcal{X}, \theta) = &\prod_{x \in \mathcal{X}} m_{d(x)}^x (1 - m_{d(x)})^{1-x} \prod_{d \in \mathcal{D}} \frac{m_d^{a_{k(d)}-1} (1-m_d)^{b_{k(d)}-1}}{\beta\left(a_{k(d)}, b_{k(d)}\right)} \cdot \\ &\prod_{l \in \mathcal{L}} \frac{{v_a}^{u_a}}{\Gamma(u_a)} a_l^{u_a - 1} e^{-v_a a_l} \frac{{v_b}^{u_b}}{\Gamma(u_b)} b_l^{u_b - 1} e^{-v_b b_l},\end{aligned} \tag{6}$$

The logarithm of it, which will prove useful immediately can be written as

$$\begin{aligned}\log \mathcal{E}(\mathcal{M},\mathcal{A},\mathcal{B}|\mathcal{X},\theta) = &\sum_{d\in\mathcal{D}}\sum_{x\in\mathcal{X}(d)} x\log m_d + (1-x)\log(1-m_d)\\ &+\sum_{d\in\mathcal{D}}(a_{k(d)}-1)\log m_d + (b_{k(d)}-1)\log(1-m_d) - \log\beta\left(a_{k(d)},b_{k(d)}\right)\\ &+\sum_{l\in\mathcal{L}} u_a\log v_a + (u_a-1)\log a_l - v_a a_l - \log\Gamma(u_a)\\ &+\sum_{l\in\mathcal{L}} u_b\log v_b + (u_b-1)\log b_l - v_b b_l - \log\Gamma(u_b)\end{aligned} \tag{7}$$

To use the Variational Bayes, we are interested in following quantities

$$\begin{aligned}\log q(m_d) =&\mathrm{E}_{\mathcal{D}\setminus d,\mathcal{A},\mathcal{B}}\left[\log\left(\mathcal{E}(\mathcal{M},\mathcal{A},\mathcal{B}|\mathcal{X},\theta)\right)\right],\\ \log q(a_l) =&\mathrm{E}_{\mathcal{D},\mathcal{A}\setminus a_l,\mathcal{B}}\left[\log\left(\mathcal{E}(\mathcal{M},\mathcal{A},\mathcal{B}|\mathcal{X},\theta)\right)\right],\\ \log q(b_l) =&\mathrm{E}_{\mathcal{D},\mathcal{A},\mathcal{B}\setminus b_l}\left[\log\left(\mathcal{E}(\mathcal{M},\mathcal{A},\mathcal{B}|\mathcal{X},\theta)\right)\right].\end{aligned} \tag{8}$$

Again the notation $\mathcal{D}\setminus d$ denotes the set of domains $\mathcal{D}$ without the element $d$, and similarly for $\mathcal{A}$ and $\mathcal{B}$. Note that for brevity, all terms not depending on the variable omitted from the expectation will be skipped, as they will become part of the normalizing constant making $q(\cdot)$ a proper probability distribution. We therefore use $\propto$ instead of $=$ to signalized the equality up to multiplicative constant.

**Derivation of** $\log q(m_d)$

For the $\log q(m_d)$ holds

$$\begin{aligned}\log q(m_d) \propto &\sum_{x\in\mathcal{X}(d)} x\log m_d + (1-x)\log(1-m_d)\\ &+(a_{k(d)}-1)\log m_d + (b_{k(d)}-1)\log(1-m_d)\\ \propto &\log m_d\left(\sum_{x\in\mathcal{X}(d)} x + a_{k(d)} - 1\right) + \log(1-m_d)\left(\sum_{x\in\mathcal{X}(d)}(1-x) + b_{k(d)} - 1\right)\end{aligned} \tag{9}$$

in which we recognize a logarithm of the beta distribution. Therefore

$$q(m_d) = \mathrm{Beta}\left(\sum_{x\in\mathcal{X}(d)} x + \widehat{a_{k(d)}},\ \sum_{x\in\mathcal{X}(d)}(1-x) + \widehat{b_{k(d)}}\right), \tag{10}$$

where $\hat{}$ denotes expected value of the variable.

**Derivation of** $q(a_k)$ **and** $q(b_k)$

Since the derivation of $q(a_k)$ and $q(b_k)$ follows exactly the same steps, it is shown only fro $q(a_k)$. For it holds

$$\begin{aligned}\log q(a_l) \propto &\mathrm{E}_{\mathcal{B},\mathcal{A}\setminus a_l}\left[\sum_{\{d\in\mathcal{D}|l\in k(d)\}} a_l a_{k(d)\setminus l}\log m_d - \log\beta(a_{a_l a_{k(d)\setminus l}}, b_{k(d)})\right.\\ &\left.-(u_a-1)\log a_l - v_a a_l,\right]\end{aligned} \tag{11}$$

where $a_{k(d)\setminus l} = \prod_{l'\in k(d)\setminus l} a_{l'}$ and likewise for $b_{k(d)\setminus l}$. It is the term $-\log\beta(a_{k(d)},b_{k(d)})$ which causes the problem, as otherwise $q(a_k)$ would have the desired form of Gamma distribution, as the posterior would be the same as the prior. A remedy is to lower bound $\log q(a_l)$ by deriving lower bound of $-\log\beta(a_{k(d)},b_{k(d)})$, instead. By maximising the new lower bound we maximize the joint pdf as well. The derivation follows the steps [2] with the difference that here $a_{k(d)} = \prod_{l\in k(d)} a_l$. It relies on the two following properties copied from [2] without the proof.

**Property 1.** *The normalization coefficient of Beta distribution can be approximated in point $x_0$ using pseudo-Taylor [2] approximation as*

$$-\log\beta(x,y) \geq -\log\beta(x_0,y) + [\psi(x_0+y)-\psi(x_0)]x_0(\log x - \log x_0), \tag{12}$$

*where the $\psi$ is Digamma function, which is defined as first derivative of a logarithm of Gamma function. This inequality holds for $y>1$, and $x, x_0 \in \mathbb{R}$.*

**Property 2.** *Digamma function can be approximated in point $y_0$ using pseudo-Taylor approximation as*

$$\psi(x_0+y) \geq \psi(x_0+y_0) + \psi'(x_0+y_0)y_0(\log y - \log y_0), \tag{13}$$

*where the $\psi'$ is Trigamma function, which is defined as second derivative of a logarithm of Gamma function. This inequality holds for $y>1$, and $x, x_0 \in \mathbb{R}$.*

The expectation of $-\log\beta(a_{k(d)}, b_{k(d)})$ can be lower-bounded as

$$\begin{aligned}\mathrm{E}_{\mathcal{B},\mathcal{A}\setminus a_l}\left[-\log\beta(a_{k(d)},b_{k(d)})\right] \underset{(1)}{\geq} &\mathrm{E}_{\mathcal{B},\mathcal{A}\setminus a_l}\left[-\log\beta(a_0,b_{k(d)}) + \right.\\ &\left. + \left[\left(\psi(a_0+b_{k(d)})-\psi(a_0)\right)a_0\left(\log a_{k(d)} - \log a_0\right)\right]\right]\\ \underset{(2)}{\propto} &\mathrm{E}_{\mathcal{B}}\left[a_0\left(\psi(a_0+b_{k(d)})-\psi(a_0)\right)\log a_l\right],\end{aligned} \tag{14}$$

where in $(1)$ we have used Property 1, and in $(2)$ we have dropped terms not depending on $a_l$ and used the fact that $\log a_{k(d)} = \sum_{l\in k(d)} \log a_l$.

To further break the term $\mathrm{E}_B\left[\psi(a_0+b_{k(d)})\right]$ we use the Property 2 as

$$\begin{aligned}\mathrm{E}_{\mathcal{B}}\left[\psi(a_0+b_{k(d)})\right] \geq &\mathrm{E}_{\mathcal{B}}\left[\psi(a_0+b_0)+\psi'(a_0+b_0)b_0(\log b_{k(d)} - \log b_0)\right]\\ = &\psi(a_0+b_0)+\psi'(a_0+b_0)b_0\left(\mathrm{E}_{\mathcal{B}}[\log b_{k(d)}] - \log b_0\right),\end{aligned} \tag{15}$$

where the approximation again used Taylor expansion in point $b_0$ and the expectation operator is moved deeper inside the brackets. Note that no terms can be dropped in (15), because $\mathrm{E}_{\mathcal{B}}\left[\psi(a_0+b_{k(d)})\right]$ is multiplicative of $\log a_l$ in (14).

By substituting (15) into (14) we obtain

$$\begin{aligned}\mathrm{E}_{\mathcal{B},\mathcal{A}\setminus a_l}\left[-\log\beta(a_{k(d)},b_{k(d)})\right] &\geq \left[\mathrm{E}_{\mathcal{B}}\left[\psi(a_0+b_{k(d)})\right]-\psi(a_0)\right]a_0\log a_l\\ &\geq \left[\psi(a_0+b_0)-\psi(a_0)+b_0\psi'(a_0+b_0)(\mathrm{E}_{\mathcal{B}}[\log b_{k(d)}]-\log b_0)\right]a_0\log a_l\\ &= \zeta_{d,a_l}\log a_l,\end{aligned} \tag{16}$$

where $\zeta_{d,a_l} = \left[\psi(a_0+b_0)-\psi(a_0)+b_0\psi'(a_0+b_0)(\mathrm{E}_{\mathcal{B}}[\log b_{k(d)}]-\log b_0)\right]a_0$.

Finally, by substituting (16) into (18) a final lower bound on $\log q(a_l)$ is obtained as

$$\begin{aligned}\log q(a_l) \propto &\mathrm{E}_{\mathcal{B},\mathcal{A}\setminus a_l}\left[a_k \sum_{\{d\in\mathcal{D}|k\in k(d)\}} a_{k(d)\setminus k}\log m_d + \zeta_{d,a_l}\log a_l\right.\\ &\left.-(u_a-1)\log a_k - v_a a_k\right],\end{aligned} \tag{17}$$

where a Gamma distribution can be recognized. Hence

$$\log q(a_l) = \mathrm{Gamma}\left(u_a + \sum_{\{d\in\mathcal{D}|k\in k(d)\}}\zeta_{d,a_l},\ v_a - \sum_{\{d\in\mathcal{D}|k\in k(d)\}}\hat{a}_{k(d)\setminus k}\widehat{\log m_d}\right), \tag{18}$$

where again $\hat{\cdot}$ is the expected value.

In the algorithm described in the paper $a_0 = \widehat{a_{k(d)}}$ and $b_0 = \widehat{b_{k(d)}}$ from the previous iteration.

## References


[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] Zhanyu Ma and Arne Leijon. Bayesian estimation of beta mixture models with variational inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2160–2173, 2011.