A The KKT Conditions

The optimization problem for variational inference on LDA is:

$$\begin{aligned} \min_{\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\phi}} \quad & KL(q(\boldsymbol{\varphi},\boldsymbol{\theta},\mathbf{Z} \mid \boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\phi}) \| p(\boldsymbol{\varphi},\boldsymbol{\theta},\mathbf{Z} \mid \mathbf{W},\alpha,\beta)) \\ \text{s.t.} \quad & -\eta_{kv} \leq 0, \quad \forall k, v; \\ & -\gamma_{dk} \leq 0, \quad \forall d, k; \\ & -\phi_{dik} \leq 0, \quad \forall d, i, k; \\ & \sum_k \phi_{dik} = 1, \quad \forall d, i. \end{aligned} \tag{20}$$

To derive the KKT conditions, we first introduce KKT multipliers $\lambda_{\eta_{kv}}, \lambda_{\gamma_{dk}}, \lambda_{\phi_{dik}}$ and $\rho_{\phi_{di}}$ to each constraint in Eq (20). The KKT conditions have four parts:

Stationarity

$$\begin{aligned} \eta_{kv} - \beta - \sum_d \sum_i \phi_{dik} \mathbb{I}_1(w_{di} = v) - \lambda_{\eta_{kv}} &= 0 \\ \gamma_{dk} - \alpha - \sum_i \phi_{dik} - \lambda_{\gamma_{dk}} &= 0 \\ \log \phi_{dik} - (\Psi(\gamma_{dk}) - \sum_{k'} \Psi(\gamma_{dk'}) + \Psi(\eta_{kw_{di}}) - \Psi(\sum_{v'} \eta_{kv})) - 1 - \lambda_{\phi_{dik}} + \rho_{\phi_{di}} &= 0. \end{aligned} \tag{21}$$

Complementary Slackness

$$\begin{aligned} -\lambda_{\eta_{kv}} \eta_{kv} &= 0 \\ -\lambda_{\gamma_{dk}} \gamma_{dk} &= 0 \\ -\lambda_{\phi_{dik}} \phi_{dik} &= 0. \end{aligned} \tag{22}$$

Primal Feasibility

$$\begin{aligned} -\eta_{kv} &\leq 0 \\ -\gamma_{dk} &\leq 0 \\ -\phi_{dik} &\leq 0 \\ \sum_k \phi_{dik} - 1 &= 0. \end{aligned} \tag{23}$$

Dual Feasibility

$$\begin{aligned} \lambda_{\eta_{kv}} &\geq 0 \\ \lambda_{\gamma_{dk}} &\geq 0 \\ \lambda_{\phi_{dik}} &\geq 0. \end{aligned} \tag{24}$$

First, we observe that Eq (21) implies that $-\eta_{kv}$, $-\gamma_{dk}$ and $-\phi_{dik}$ are both strictly negative because:

$$\begin{aligned} -\eta_{kv} \leq -\beta &< 0 \\ -\gamma_{dk} \leq -\alpha &< 0 \\ -\phi_{dik} = -\exp((\Psi(\gamma_{dk}) - \sum_{k'} \Psi(\gamma_{dk'}) + \Psi(\eta_{kw_{di}}) - \Psi(\sum_{v'} \eta_{kv})) - 1 - \lambda_{\phi_{dik}} + \rho_{\phi_{di}}) &< 0. \end{aligned}$$

We combine the above result with the complementary slackness Eq (22):

$$\begin{aligned} \lambda_{\eta_{kv}} &= 0 \\ \lambda_{\gamma_{dk}} &= 0 \\ \lambda_{\phi_{dik}} &= 0. \end{aligned} \tag{25}$$

We plug Eq (25) into Eqs (21) and (23):

$$\begin{aligned}
\eta_{kv} - \beta - \sum_d \sum_i \phi_{dik} \mathbb{I}_1(w_{di} = v) &= 0 \\
\gamma_{dk} - \alpha - \sum_i \phi_{dik} &= 0 \\
\log \phi_{dik} - (\Psi(\gamma_{dk}) - \sum_{k'} \Psi(\gamma_{dk'}) + \Psi(\eta_{kw_{di}}) - \Psi(\sum_{v'} \eta_{kv})) - 1 + \rho_{\phi_{di}} &= 0 \\
\rho_{\phi_{di}} - \log[\sum_k \exp(\Psi(\gamma_{dk}) - \sum_{k'} \Psi(\gamma_{dk'}) + \Psi(\eta_{kw_{di}}) - \Psi(\sum_{v'} \eta_{kv}) + 1)] &= 0. \qquad (26)
\end{aligned}$$

Eqs (26) and (25) are equivalent with the KKT conditions in Eqs (21),(23),(24) and (22). We focus on the conditions on primal variables and further simplify Eq (26) to get the equivalent form of KKT condition:

$$\begin{aligned}
\eta_{kv} - \beta - \sum_d \sum_i \phi_{dik} \mathbb{I}_1(w_{di} = v) &= 0 \\
\gamma_{dk} - \alpha - \sum_i \phi_{dik} &= 0 \\
\phi_{dik} - \frac{\exp(\Psi(\gamma_{dk}) + (\Psi(\eta_{kw_{di}}) - \Psi(\sum_{v'} \eta_{kv})))}{\sum_k \exp(\Psi(\gamma_{dk}) + (\Psi(\eta_{kw_{di}}) - \Psi(\sum_{v'} \eta_{kv'})))} &= 0. \qquad (27)
\end{aligned}$$

These are exactly the variational inference formulas in (Blei et al. 2003). After changing the notation (discussed in the main paper), we get Eq (4).

B Implicit Functions

We review the definition of implicit functions. We denote the $\epsilon-$ball of $\mathbf{x} \in \mathbb{R}^d$ as $N(\mathbf{x}, \epsilon) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\|_2 < \epsilon\}$. We call $\hat{\mathbf{y}}(\mathbf{x}) \in \mathbb{R}^m$ an implicit function of $\mathbf{x} \in \mathbb{R}^n$ defined by implicit equation $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$, where $\mathbf{f}(\mathbf{x}, \mathbf{y}) : \mathbb{R}^{n+m} \mapsto \mathbb{R}^m$, if for any $\mathbf{x}$, $\hat{\mathbf{y}}(\mathbf{x})$ satisfies $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ and there exists $\epsilon > 0$ such that for any $\mathbf{x} + \delta \in N(\mathbf{x}, \epsilon)$, only $\hat{\mathbf{y}}(\mathbf{x} + \delta)$ both satisfies $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ and belongs to $N(\hat{\mathbf{y}}(\mathbf{x}), \epsilon)$.

Then we show $\hat{\boldsymbol{\mu}}(\mathbf{M})$ is an implicit function of $\boldsymbol{\mu}$ defined by implicit equations $\mathbf{f}_{\boldsymbol{\mu}} = \mathbf{0}$. First, $\mathbf{f}_{\boldsymbol{\mu}}$ is a (multivariate) continuous differential function of $\boldsymbol{\mu}$ because each component of $\mathbf{f}_{\boldsymbol{\mu}}$ consists of continuous differential terms. Second, by the assumption in the theorem, $\mathbf{f}_{\boldsymbol{\mu}}$ has an invertible Jacobian matrix $\frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \boldsymbol{\mu}}$. Therefore, according to implicit function theorem (Danilov 2001), $\hat{\boldsymbol{\mu}}(\mathbf{M})$ is an implicit function of $\mathbf{M}$ and $\hat{\boldsymbol{\mu}}(\mathbf{M})$ is continuously differentiable with respect to $\mathbf{M}$. The gradient is as in Eq (14).

C Fast Approximation for Computing $\frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}}$

Theorem 1 *If $\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \boldsymbol{\phi}} = \mathbf{0}$, the element at kv-th row and dv′-th column in the Jacobian matrix $\frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}}$ is*

$$[\frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}}]_{kv,dv'} = \phi_{dvk} \mathbb{I}_1(v = v'). \qquad (28)$$

Proof: First, $\frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}}$ is the first $N_{\boldsymbol{\eta}}$ rows of the size $(N_{\boldsymbol{\eta}} + N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}}) \times N_{\mathbf{M}}$ Jacobian matrix $\frac{\partial \hat{\boldsymbol{\mu}}(\mathbf{M})}{\partial \mathbf{M}}$,

$$\frac{\partial \hat{\boldsymbol{\mu}}(\mathbf{M})}{\partial \mathbf{M}} = \begin{bmatrix} \frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}} \\ \frac{\partial (\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}})(\mathbf{M})}{\partial \mathbf{M}} \end{bmatrix},$$

where $\frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}}$ is a $N_{\boldsymbol{\eta}} \times N_{\mathbf{M}}$ Jacobian matrix, $\frac{\partial (\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}})(\mathbf{M})}{\partial \mathbf{M}}$ is a $(N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}}) \times N_{\mathbf{M}}$ Jacobian matrix.

We define a selection matrix $\mathbf{P} \triangleq [\mathbf{I}\ \mathbf{0}]$ (size of $N_{\boldsymbol{\eta}} \times (N_{\boldsymbol{\eta}} + N_{\mathbf{M}} + N_{\boldsymbol{\phi}})$) and $\frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}}$ is selected from $\frac{\partial \hat{\boldsymbol{\mu}}(\mathbf{M})}{\partial \mathbf{M}}$ by multiplying $\mathbf{P}$ on the left:

$$\frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}} = \mathbf{P} \frac{\partial \hat{\boldsymbol{\mu}}(\mathbf{M})}{\partial \mathbf{M}}. \qquad (29)$$

$\frac{\partial \hat{\boldsymbol{\mu}}(\mathbf{M})}{\partial \mathbf{M}}$ is computed as in Eq (14). We introduce the two terms on the right side of Eq (14) respectively.

The first term, $(\frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \boldsymbol{\mu}})^{-1}$ is the inversion of a Jacobian matrix with size of $(N_{\boldsymbol{\eta}} + N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}}) \times (N_{\boldsymbol{\eta}} + N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}})$. Similar to the divide of $\frac{\partial \hat{\boldsymbol{\mu}}(\mathbf{M})}{\partial \mathbf{M}}$, we write $\frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \boldsymbol{\mu}}$ (and correspondingly its inversion) as 4 blocks:

$$(\frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \boldsymbol{\mu}})^{-1} = \begin{bmatrix} \frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \boldsymbol{\eta}} & \frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial (\boldsymbol{\gamma}, \boldsymbol{\phi})} \\ \frac{\partial (\mathbf{f}_{\boldsymbol{\gamma}}, \mathbf{f}_{\boldsymbol{\phi}})}{\partial \boldsymbol{\eta}} & \frac{\partial (\mathbf{f}_{\boldsymbol{\gamma}}, \mathbf{f}_{\boldsymbol{\phi}})}{\partial (\boldsymbol{\gamma}, \boldsymbol{\phi})} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}, \tag{30}$$

where $\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \boldsymbol{\eta}}$ and $\mathbf{A}$ have size $N_{\boldsymbol{\eta}} \times N_{\boldsymbol{\eta}}$, $\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial (\boldsymbol{\gamma}, \boldsymbol{\phi})}$ and $\mathbf{B}$ have size $N_{\boldsymbol{\eta}} \times (N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}})$, $\frac{\partial (\mathbf{f}_{\boldsymbol{\gamma}}, \mathbf{f}_{\boldsymbol{\phi}})}{\partial \boldsymbol{\eta}}$ and $\mathbf{C}$ have size $(N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}}) \times N_{\boldsymbol{\eta}}$, and $\frac{\partial (\mathbf{f}_{\boldsymbol{\gamma}}, \mathbf{f}_{\boldsymbol{\phi}})}{\partial (\boldsymbol{\gamma}, \boldsymbol{\phi})}$ and $\mathbf{D}$ have size $(N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}}) \times (N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}})$.

The second term on the right side $\frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \mathbf{M}}$ has $(N_{\boldsymbol{\eta}} + N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}})$ rows and $N_{\mathbf{M}}$ columns. We write it as two blocks according to the division of $\frac{\partial \hat{\boldsymbol{\mu}}(\mathbf{M})}{\partial \mathbf{M}}$,

$$\frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \mathbf{M}} = \begin{bmatrix} \frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \mathbf{M}} \\ \frac{\partial (\mathbf{f}_{\boldsymbol{\gamma}}, \mathbf{f}_{\boldsymbol{\phi}})}{\partial \mathbf{M}} \end{bmatrix}, \tag{31}$$

where $\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \mathbf{M}}$ has size $N_{\boldsymbol{\eta}} \times N_{\mathbf{M}}$ and $\frac{\partial (\mathbf{f}_{\boldsymbol{\gamma}}, \mathbf{f}_{\boldsymbol{\phi}})}{\partial \mathbf{M}}$ has size $(N_{\boldsymbol{\gamma}} + N_{\boldsymbol{\phi}}) \times N_{\mathbf{M}}$.

We plug the block form of matrices in Eqs (29), (30) and (31) into Eq (14):

$$\frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}} = -\mathbf{P} \left(\frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \boldsymbol{\mu}} \right)^{-1} \frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \mathbf{M}} = -\begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \mathbf{M}} \\ \frac{\partial (\mathbf{f}_{\boldsymbol{\gamma}}, \mathbf{f}_{\boldsymbol{\phi}})}{\partial \mathbf{M}} \end{bmatrix} = -\left(\mathbf{A} \frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \mathbf{M}} + \mathbf{B} \frac{\partial (\mathbf{f}_{\boldsymbol{\gamma}}, \mathbf{f}_{\boldsymbol{\phi}})}{\partial \mathbf{M}} \right). \tag{32}$$

Now we need to calculate the two blocks $\mathbf{A}$ and $\mathbf{B}$ of the inverted Jacobian matrix $(\frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \boldsymbol{\mu}})^{-1}$. In the assumption of theorem, $\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \boldsymbol{\phi}} = \mathbf{0}$. Note that $\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \boldsymbol{\phi}} = \mathbf{0}$. According to Eq (2), we have $\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \boldsymbol{\gamma}} = \mathbf{0}$ and $\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial (\boldsymbol{\gamma}, \boldsymbol{\phi})} = \mathbf{0}$. Therefore, $\frac{\partial \mathbf{f}_{\boldsymbol{\mu}}}{\partial \boldsymbol{\mu}}$ is a blockwise lower triangle matrix. Based on the property of blockwise inversion of matrix, we get

$$\mathbf{A} = \mathbf{I}, \mathbf{B} = \mathbf{0}.$$

We put the values of $\mathbf{A}$ and $\mathbf{B}$ into Eq (32) and get

$$\frac{\partial \hat{\boldsymbol{\eta}}(\mathbf{M})}{\partial \mathbf{M}} = -\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \mathbf{M}}, \tag{33}$$

where each element in $\frac{\partial \mathbf{f}_{\boldsymbol{\eta}}}{\partial \mathbf{M}}$ is calculated by Eq (2):

$$\frac{\partial f_{\eta_{kv}}(\boldsymbol{\mu}, \mathbf{M})}{\partial m_{dv'}} = -\phi_{dvk} \mathbb{I}_1(v = v'). \tag{34}$$

We combine Eq (33) and Eq (34) and get

$$\nabla_{m_{dv'}} \hat{\eta}_{kv}(\mathbf{M}) = -(-\phi_{dvk} \mathbb{I}_1(v = v')) = \phi_{dvk} \mathbb{I}_1(v = v'). \tag{35}$$

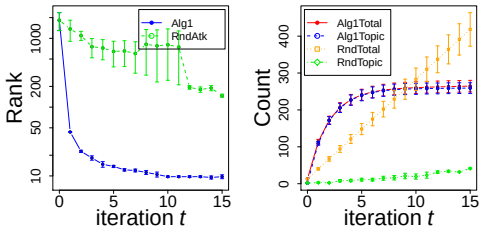
■

We note that in practice Theorem 1's condition does not hold. Nonetheless, Theorem 1 provides an approximation to $\nabla_{\mathbf{M}} \hat{\eta}_{kv}(\mathbf{M})$. In our experiments, this approximation works well.

D Detailed Experiment Results

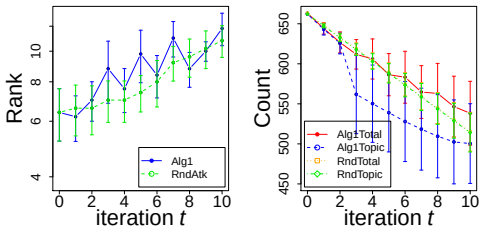
D.1 Rank and Contribution of "ceiling" in Promote-Word Attack

The rank and contribution of word "ceiling" in promote-word attack on AP are shown in Figure 5. The results are very similar as results of "marijuana" and "debt" in promote-word attack shown in main paper.



(a) ceiling rank (b) ceiling contribution

Figure 5: (second-part of) Promote-word attack on word "debt" and "ceiling" in the *market* topic from AP.



(a) Paul rank (b) Paul contribution

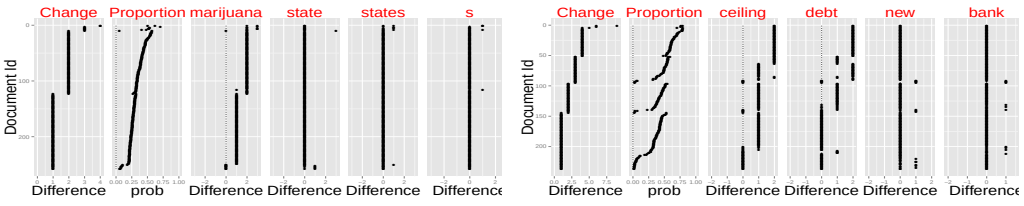
Figure 6: (second-part of) Replace-word attack to replace word "Paul" with "Weasley" in the *president* topic from WISH.

D.2 Rank and Contribution of "Paul" in Replace-Word Attack

We show the rank and contribution of word "Paul" in replace-word attack on WISH in Figure 6. The results are very similar as results of "Iraq" in demote-word attack shown in main paper.

D.3 Detailed Attack Behavior of Promote-Word Attacks

We show the detailed attack behavior of Alg1 in promote-word attack on CONG and AP on documents in Figure 7. The documents are shown from up to down sorted by the decreasing amount of changes defined in Eq (19). For each document d, we show the target topic proportion $\hat{\theta}_{dk} \triangleq \gamma_{dk} / \sum_k \gamma_{dk}$ and the count changes (in the document) of 4 words which have the largest changes on the whole corpus.

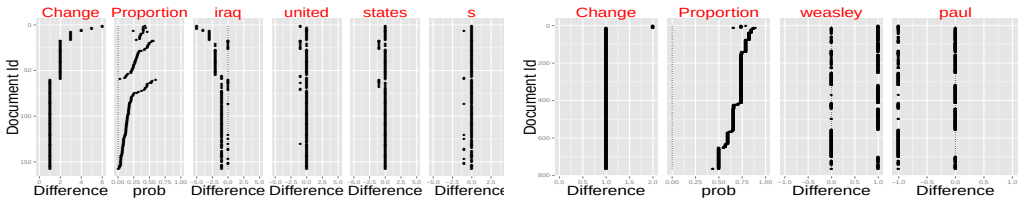


(a) Promote "marijuana" Attack (b) Promote "debt" and "ceiling" Attack

Figure 7: Statistics of attack behavior of promote-word attacks

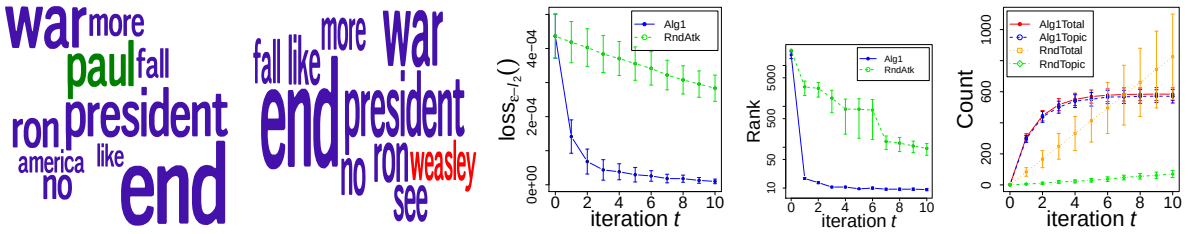
D.4 Detailed Attack Behavior of Demote-Word Attack

We plot the attack behavior of the Alg1 in demote-word attack on CONG in Figure 8(a). Things we show are exactly the same as in promote-word attack.



(a) Demote "Iraq" Attack (b) Replace "Paul" with "Weasley" Attack

Figure 8: Statistics of attack behavior of demote "Iraq" attack and replace "Paul" with "Weasley" attack



(a) before attack (b) after attack (c) objective (d) Weasley rank (e) Weasley contribution

Figure 9: Replace-word attack on "Paul" and "Weasley" in the *president* topic from WISH

D.5 Replace-Word Attack

This kind of attack replaces a top-10 word in the target topic with another word. LDA on the original WISH corpus consistently produced a *president* topic with "Paul" (as in Ron Paul) as a top word. To demonstrate replace-word attack, we replace "Paul" with "Weasley" (as in Ron Weasley of Harry Potter fame). The target encoding is a combination of promotion and demotion with $\varphi^*_{k,\text{paul}} = \varphi_{k,w_{11}}$ and $\varphi^*_{k,\text{weasley}} = \varphi_{k,w_9}$, then renormalize φ^*_k. The RndAtk baseline is also a combination which randomly adds "Weasley" and deletes "Paul", subject to the constraint encoded in $\mathbb{M}$. Similar to previous attacks, Figure 9 shows Alg1's effectiveness in the replace-word attack. In Panel (a,b) Alg1 successfully replaced word "Paul" with "Weasley" in top-10 words. "Paul" ranked 11th after attack. The objective function are optimized rapidly in Panel (c). The rank and contribution of "Weasley" in Panel (d,e) are similar to the promote-word attack, and those of "Paul" are similar to the demote-word attack. Alg1's attack behavior was a combination of promote-word attack and demote-word attack. It mainly replaced "Paul" with "Weasley" in selected documents with high target topic proportion. We plot the attack behavior of the Alg1 in replace-word attack on WISH in Figure 8(b). The settings of things we show are exactly the same as in previous attacks. Modification on only two words "Paul" and "Weasley" is shown because no modification exists on other words.

D.6 Detailed Attack Behavior of Attack with POS Constraint

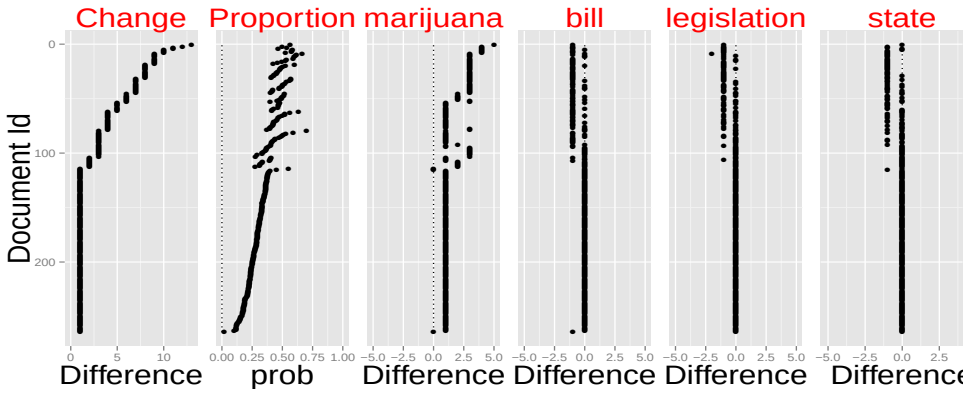


Figure 10: Statistics of attack behavior of attack with POS constraint

We plot the attack behavior of the Alg1 in promote-word attack with POS constraint on CONG in Figure 10. The things we show are exactly the same as in previous attacks.

D.7 Detailed Attack Behavior of Attack with POS Constraint

We plot the attack behavior of the Alg1 in sentence attack on WISH in Figure 11. The things we show are exactly the same as in previous attacks.

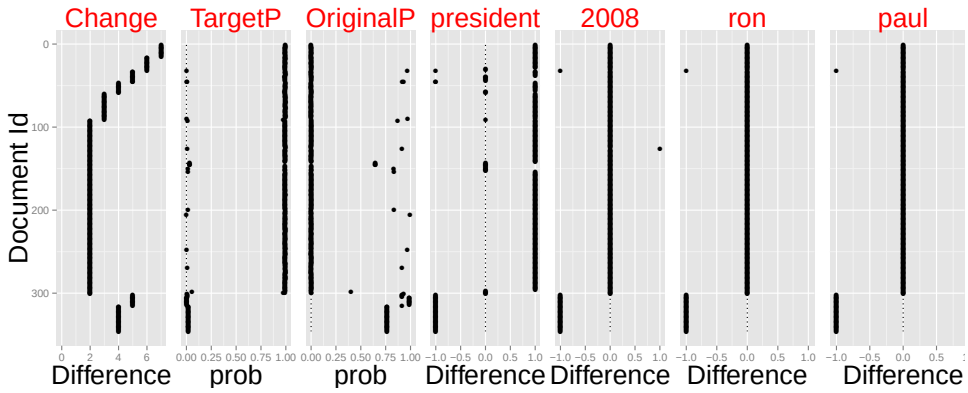


Figure 11: Statistics of attack behavior of attack with sentence attack