## Supplementary A: Formulation of (6)

$p(a_\tau|s_\tau)$ in Equation (6) is entangled by actions for exploring edge existence and influence, and the probability to choose influence is deterministic and non-differentiable. In order to back-propagate gradients, we approximate the optimization problem. It is apparent the existence and influence are independent, so we rewrite the gradient:

$$\nabla_\theta Z = -\mathbb{E}_\pi[r_\tau(s_\tau, a_\tau)(\nabla_\theta \log p(a_\tau^e|s_\tau) + \nabla_\theta \log p(a_\tau^i|s_\tau))]$$

where $a_\tau^e$ means the action to change existence of the graph i.e. creation, deletion and unchange edges and $a_\tau^i$ represents the choice of parameters for influence i.e. the value of $K_{ts}^l$.

As for the former transition probability of existence, we denote $-\mathbb{E}_\pi[r_\tau(s_\tau, a_\tau)\nabla_\theta \log p(a_\tau^e|s_\tau)]$ as $L^e$ and rewrite it as follows:

$$\begin{aligned}
L^e &= -\mathbb{E}_\pi[r_\tau(s_\tau, a_\tau)((\nabla_\theta \log p(a_\tau^{e,c}|s_\tau) + \\
&\quad \nabla_\theta \log p(a_\tau^{e,d}|s_\tau, a_\tau^{e,c}) + \nabla_\theta \log p(a_\tau^{e,u}|s_\tau, a_\tau^{e,c}, a_\tau^{e,d}))] \\
&= -\mathbb{E}_\pi[r_\tau(s_\tau, a_\tau)((\frac{\partial logp(a_\tau^{e,c}|s_\tau)}{\partial W_{et}^l}\frac{\partial W_{et}^l}{\partial \theta} + \\
&\quad \frac{\partial logp(a_\tau^{e,d}|s_\tau, a_\tau^{e,c})}{\partial W_{et}^{'l}}\frac{\partial W_{et}^{'l}}{\partial W_{et}^l}\frac{\partial W_{et}^l}{\partial \theta} + 0))] \\
&= -\mathbb{E}_\pi[r_\tau(s_\tau, a_\tau)((\frac{\partial logp(a_\tau^{e,c}|s_\tau)}{\partial W_{et}^l}\frac{\partial W_{et}^l}{\partial \theta} + \\
&\quad \alpha\frac{\partial logp(a_\tau^{e,d}|s_\tau, a_\tau^{e,c})}{\partial W_{et}^{'l}} inv(W_{et}^l)\frac{\partial W_{et}^l}{\partial \theta}))] \qquad (1)
\end{aligned}$$

where $inv(W)$ means a matrix whose elements are reciprocal with those in $W$ and $\alpha$ is a small hyperparameter to approximate the derivative of normalizing $W_{et}^{'l}$ because it is computationally expensive.

Since the sampling strategy for influence is deterministic, we apply a symmetric clip function to approximate the integral of delta distribution. At first, we explicitly write the probability of actions for influence selection:

$$p(a_\tau^i|s_\tau) = \begin{cases} \int_{-\frac{|a_\tau^i|-1}{2K_0}}^{-\frac{|a_\tau^i|-3}{2K_0}} \delta(t - W_{it}^l)dt & a_\tau^i < 0 \\ \int_{\frac{|a_\tau^i|-3}{2K_0}}^{\frac{|a_\tau^i|-1}{2K_0}} \delta(t - W_{it}^l)dt & a_\tau^i > 0 \end{cases} \qquad (2)$$

It is apparent that $\log p(a_\tau^i|s_\tau)$ is non-differentiable of $W_{it}^l$. For any element $w_{it}^l \in W_{it}^l$, we use the following function to substitute original objective:

$$p(a_\tau^i|s_\tau) \approx \begin{cases} 2K_0 w_{it}^l - a_\tau^i + 3 & w_{it}^l \in (\frac{a_\tau^i - 3}{2K_0}, \frac{a_\tau^i - 2}{2K_0}] \cup w_{it}^l \geqslant 0 \\ -2K_0 w_{it}^l + a_\tau^i - 1 & w_{it}^l \in (\frac{a_\tau^i - 2}{2K_0}, \frac{a_\tau^i - 1}{2K_0}] \cup w_{it}^l \geqslant 0 \\ 2K_0 w_{it}^l - a_\tau^i - 1 & w_{it}^l \in (\frac{a_\tau^i + 1}{2K_0}, \frac{a_\tau^i + 2}{2K_0}] \cup w_{it}^l < 0 \\ -2K_0 w_{it}^l + a_\tau^i + 3 & w_{it}^l \in (\frac{a_\tau^i + 2}{2K_0}, \frac{a_\tau^i + 3}{2K_0}] \cup w_{it}^l < 0 \end{cases} \qquad (3)$$

We can rewrite the gradient for updating the parameters in the policy network to learn the influence in REINFORCE algorithm. We use $L_i$ to substitute $-\mathbb{E}_\pi[r_\tau(s_\tau, a_\tau)\nabla_\theta \log p(a_\tau^i|s_\tau)]$ as follows:

$$L_i = -\mathbb{E}_\pi[r_\tau(s_\tau, a_\tau) \cdot 2K_0 \cdot \frac{w_{it}^l - \frac{|a_\tau^i|-2}{2K_0}\frac{a_\tau^i}{|a_\tau^i|}}{|w_{it}^l - \frac{|a_\tau^i|-2}{2K_0}\frac{a_\tau^i}{|a_\tau^i|}|}] \qquad (4)$$

where the logarithm in the original gradient is removed for simplicity.

In summary, the approximation of the expected gradient can be written as follows:

$$\begin{aligned}
\nabla_\theta Z_{approx} = -\frac{1}{M}\sum_{k=1}^{M} R_k \sum_{\tau=0}^{T_k}(&\frac{\partial logp(a_\tau^{e,c}|s_\tau)}{\partial W_{et}^l}\frac{\partial W_{et}^l}{\partial \theta} + \\
&\alpha\frac{\partial logp(a_\tau^{e,d}|s_\tau, a_\tau^{e,c})}{\partial W_{et}^{'l}} inv(W_{et}^l)\frac{\partial W_{et}^l}{\partial \theta} + 2K_0 \cdot \frac{W_{it}^l - \frac{|a_\tau^i|-2}{2K_0}\frac{a_\tau^i}{|a_\tau^i|}}{|W_{it}^l - \frac{|a_\tau^i|-2}{2K_0}\frac{a_\tau^i}{|a_\tau^i|}|})
\end{aligned} \qquad (5)$$