APPENDIX

A. DEGENERATION

In this appendix, we show that if the second-order transition probability is the same as the first-order transition probability, i.e., if $p_{i,j,k}=p_{j,k}$, each developed second-order measure degenerates to its original first-order form.

LEMMA 4. *If $p_{i,j,k}=p_{j,k}$, we have that* $\mathbf{M}=\mathbf{E}\mathbf{H}$.

PROOF. Let $u=(i,j)$ and $v=(j,k)$. There is only one non-zero element in the row vector $[\mathbf{E}]_{u,:}$, i.e., $[\mathbf{E}]_{u,j}=1$. There is only one non-zero element in the column vector $[\mathbf{H}]_{:,v}$, i.e., $[\mathbf{H}]_{j,v}=p_{j,k}$. Thus, we have $p_{j,k}=[\mathbf{E}]_{u,j}\cdot[\mathbf{H}]_{j,v}=[\mathbf{E}]_{u,:}\cdot[\mathbf{H}]_{:,v}$. We also have that $p_{i,j,k}=[\mathbf{M}]_{u,v}$. Since $p_{i,j,k}=p_{j,k}$, we have that $[\mathbf{M}]_{u,v}=[\mathbf{E}]_{u,:}\cdot[\mathbf{H}]_{:,v}$ for any two edges u and v. Thus, we have that $\mathbf{M}=\mathbf{E}\mathbf{H}$. □

THEOREM 9. *If $p_{i,j,k}=p_{j,k}$, the second-order random walk degenerates to the first-order random walk.*

PROOF. In the first-order random walk, the recursive equation is

$$\mathbf{r}=\mathbf{P}^{\mathsf{T}}\mathbf{r}=\mathbf{E}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{r}$$

Multiplying $\mathbf{H}^{\mathsf{T}}$ from left to both sides, we have that

$$\mathbf{H}^{\mathsf{T}}\mathbf{r}=\mathbf{H}^{\mathsf{T}}\mathbf{E}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{r}$$

By Lemma 4, if $p_{i,j,k}=p_{j,k}$, we have that $\mathbf{M}=\mathbf{E}\mathbf{H}$. Let $\mathbf{s}=\mathbf{H}^{\mathsf{T}}\mathbf{r}$. We have that

$$\mathbf{s}=\mathbf{H}^{\mathsf{T}}\mathbf{E}^{\mathsf{T}}\mathbf{s}=\mathbf{M}^{\mathsf{T}}\mathbf{s} \quad \text{and} \quad \mathbf{r}=\mathbf{E}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{r}=\mathbf{E}^{\mathsf{T}}\mathbf{s}$$

Thus, we get the equations for the second-order random walk. That is, the solution to the first-order random walk is also a solution to the second-order random walk. Since the solutions are unique, we can complete the proof. □

THEOREM 10. *If $p_{i,j,k}=p_{j,k}$, the second-order PageRank degenerates to the first-order PageRank.*

PROOF. In the first-order PageRank, the recursive equation is

$$\mathbf{r}=c\mathbf{P}^{\mathsf{T}}\mathbf{r}+(1-c)\mathbf{1}/n=c\mathbf{E}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{r}+(1-c)\mathbf{1}/n$$

Multiplying $\mathbf{H}^{\mathsf{T}}$ from left to both sides, we have that

$$\mathbf{H}^{\mathsf{T}}\mathbf{r}=c\mathbf{H}^{\mathsf{T}}\mathbf{E}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{r}+(1-c)\mathbf{H}^{\mathsf{T}}\mathbf{1}/n$$

By Lemma 4, if $p_{i,j,k}=p_{j,k}$, we have that $\mathbf{M}=\mathbf{E}\mathbf{H}$. Let $\mathbf{s}=\mathbf{H}^{\mathsf{T}}\mathbf{r}$. We have that

$$\mathbf{s}=c\mathbf{H}^{\mathsf{T}}\mathbf{E}^{\mathsf{T}}\mathbf{s}+(1-c)\mathbf{H}^{\mathsf{T}}\mathbf{1}/n=c\mathbf{M}^{\mathsf{T}}\mathbf{s}+(1-c)\mathbf{H}^{\mathsf{T}}\mathbf{1}/n$$

and $\mathbf{r}=c\mathbf{E}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{r}+(1-c)\mathbf{1}/n=c\mathbf{E}^{\mathsf{T}}\mathbf{s}+(1-c)\mathbf{1}/n$

Thus, we get the equations for the second-order PageRank. That is, the solution to the first-order PageRank is also a solution to the second-order PageRank. Since the solutions are unique, we can complete the proof. □

THEOREM 11. *If $p_{i,j,k}=p_{j,k}$, the second-order random walk with restart degenerates to the first-order random walk with restart.*

PROOF. The proof is similar to that of Theorem 10. □

THEOREM 12. *If $p_{i,j,k}=p_{j,k}$, the second-order SimRank degenerates to the first-order SimRank.*

PROOF. In the first-order SimRank, we have that

$$r_{i,j}=(1-c)\sum_{t=0}^{\infty}c^t\mathbb{P}[\Phi_{i,j}^{t,2t}]$$

In the second-order SimRank, we have that

$$r_{i,j}=(1-c)\sum_{t=0}^{\infty}c^t\mathbb{M}[\Phi_{i,j}^{t,2t}]$$

By Lemma 3, if $p_{i,j,k}=p_{j,k}$, we have that $\mathbb{M}[\Phi_{i,j}^{t,2t}]=\mathbb{P}[\Phi_{i,j}^{t,2t}]$. This completes the proof. □

THEOREM 13. *If $p_{i,j,k}=p_{j,k}$, the second-order SimRank* degenerates to the first-order SimRank*.*

PROOF. In the first-order SimRank*, we have that

$$r_{i,j}=(1-c)\sum_{t=0}^{\infty}\frac{c^t}{2^t}\sum_{a=0}^{t}\binom{t}{a}\mathbb{P}[\Phi_{i,j}^{a,t}]$$

In the second-order SimRank*, we have that

$$r_{i,j}=(1-c)\sum_{t=0}^{\infty}\frac{c^t}{2^t}\sum_{a=0}^{t}\binom{t}{a}\mathbb{M}[\Phi_{i,j}^{a,t}]$$

By Lemma 3, if $p_{i,j,k}=p_{j,k}$, we have that $\mathbb{M}[\Phi_{i,j}^{a,t}]=\mathbb{P}[\Phi_{i,j}^{a,t}]$. This completes the proof. □

B. VISITING PROBABILITY

The proof of Lemma 2 is as follows.

PROOF. We prove each of the four cases individually including $0=a=b$, $0<a=b$, $0=a<b$, and $0<a<b$.

In the first case, the lemma trivially holds. The probability of visiting a meeting path of length $\{0,0\}$ between nodes i and j is 1 if $i=j$ and 0 if $i\neq j$, i.e., $\mathbb{M}[\Phi_{i,j}^{0,0}]=\mathbf{I}_{i,j}$.

In the second case, we have that $0<a=b$. We proceed by induction on a. If $a=1$, we have that

$$[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}]_{x,j}=[\mathbf{H}\mathbf{E}]_{x,j}=[\mathbf{P}]_{x,j}=p_{x,j}$$

Since there is a unique path of length 1 from node x to j, which is the directed edge (x,j), we have that $\mathbb{M}[\Phi_{x,j}^{1,1}]=p_{x,j}$. Therefore, we have $\mathbb{M}[\Phi_{x,j}^{1,1}]=[\mathbf{H}\mathbf{E}]_{x,j}$ thus the lemma holds for $a=1$. Now assume that the lemma holds for $a\,(a\geq 1)$. By the assumption, we have

$$\begin{aligned}\mathbb{M}[\Phi_{x,j}^{a,a}]=[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}]_{x,j}&=\textstyle\sum_{i\in I_j}[\mathbf{H}\mathbf{M}^{a-1}]_{x,(i,j)}\cdot[\mathbf{E}]_{(i,j),j}\\&=\textstyle\sum_{i\in I_j}[\mathbf{H}\mathbf{M}^{a-1}]_{x,(i,j)}\end{aligned}$$

Each term $[\mathbf{H}\mathbf{M}^{a-1}]_{x,(i,j)}$ represents the sum of probabilities of visiting the paths of length a from node x to j whose last edge is (i,j). Next, we prove that the lemma holds for $(a+1)$. Each term $[\mathbf{H}\mathbf{M}^{a}\mathbf{E}]_{x,k}$ can be expanded as

$$\begin{aligned}[\mathbf{H}\mathbf{M}^{a}\mathbf{E}]_{x,k}&=\textstyle\sum_{j\in I_k}[\mathbf{H}\mathbf{M}^{a}]_{x,(j,k)}\cdot[\mathbf{E}]_{(j,k),k}\\&=\textstyle\sum_{j\in I_k}[\mathbf{H}\mathbf{M}^{a}]_{x,(j,k)}\\&=\textstyle\sum_{j\in I_k}\sum_{i\in I_j}[\mathbf{H}\mathbf{M}^{a-1}]_{x,(i,j)}\cdot[\mathbf{M}]_{(i,j),(j,k)}\\&=\textstyle\sum_{j\in I_k}\sum_{i\in I_j}[\mathbf{H}\mathbf{M}^{a-1}]_{x,(i,j)}\cdot p_{i,j,k}\end{aligned}$$

Consider a path ρ of length $(a+1)$ from node x to k whose last two edges are (i,j) and (j,k). The path ρ consists of a path ρ' of length a from x to j whose last edge is (i,j), followed by the edge (j,k). The probability of visiting ρ equals the probability of visiting path ρ' times the transition probability $p_{i,j,k}$. It follows that $[\mathbf{H}\mathbf{M}^{a-1}]_{x,(i,j)}\cdot p_{i,j,k}$ equals the sum of probabilities of visiting the paths of length $(a+1)$ from node x to k whose last two edges are (i,j) and (j,k). Thus, $\sum_{j\in I_k}\sum_{i\in I_j}[\mathbf{H}\mathbf{M}^{a-1}]_{x,(i,j)}\cdot p_{i,j,k}$ is the sum of probabilities of visiting all paths of length $(a+1)$ from x to k. Therefore, we have that $\mathbb{M}[\Phi_{x,k}^{a+1,a+1}]=[\mathbf{H}\mathbf{M}^{a}\mathbf{E}]_{x,k}$. This completes the proof for the second case.

In the third case, we have that $0=a<b$. We proceed by induction on b. If $b=1$, we have that

$$[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{j,y}=[\mathbf{E}^\mathsf{T}\mathbf{H}^\mathsf{T}]_{j,y}=[\mathbf{P}^\mathsf{T}]_{j,y}=p_{y,j}$$

Since there is a unique path of length 1 from node y to j, which is the directed edge (y,j), we have that $\mathbb{M}[\Phi^{0,1}_{j,y}]=p_{y,j}$. Therefore, we have $\mathbb{M}[\Phi^{0,1}_{j,y}]=[\mathbf{E}^\mathsf{T}\mathbf{H}^\mathsf{T}]_{j,y}$ thus the lemma holds for $b=1$. Now assume that the lemma holds for $b\,(b\geq 1)$. By the assumption, we have

$$\begin{aligned}\mathbb{M}[\Phi^{0,b}_{j,y}]=[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{j,y}&=\textstyle\sum_{i\in I_j}[\mathbf{E}^\mathsf{T}]_{j,(i,j)}\cdot[(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{(i,j),y}\\&=\textstyle\sum_{i\in I_j}[(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{(i,j),y}\end{aligned}$$

Each term $[(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{(i,j),y}$ represents the sum of probabilities of visiting the paths of length b from node y to j whose last edge is (i,j). Next, we prove that the lemma holds for $(b+1)$. Each term $[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b}\mathbf{H}^\mathsf{T}]_{k,y}$ can be expanded as

$$\begin{aligned}[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b}\mathbf{H}^\mathsf{T}]_{k,y}&=\textstyle\sum_{j\in I_k}[\mathbf{E}^\mathsf{T}]_{k,(j,k)}\cdot[(\mathbf{M}^\mathsf{T})^{b}\mathbf{H}^\mathsf{T}]_{(j,k),y}\\&=\textstyle\sum_{j\in I_k}[(\mathbf{M}^\mathsf{T})^{b}\mathbf{H}^\mathsf{T}]_{(j,k),y}\\&=\textstyle\sum_{j\in I_k}\sum_{i\in I_j}[\mathbf{M}^\mathsf{T}]_{(j,k),(i,j)}\cdot[(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{(i,j),y}\\&=\textstyle\sum_{j\in I_k}\sum_{i\in I_j}p_{i,j,k}\cdot[(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{(i,j),y}\end{aligned}$$

Consider a path ρ of length $(b+1)$ from node y to k whose last two edges are (i,j) and (j,k). The path ρ consists of a path ρ' of length b from y to j whose last edge is (i,j), followed by the edge (j,k). The probability of visiting ρ equals the probability of visiting path ρ' times the transition probability $p_{i,j,k}$. It follows that $p_{i,j,k}\cdot[(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{(i,j),y}$ equals the sum of probabilities of visiting the paths of length $(b+1)$ from node y to k whose last two edges are (i,j) and (j,k). Thus, $\sum_{j\in I_k}\sum_{i\in I_j}p_{i,j,k}\cdot[(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{(i,j),y}$ is the sum of probabilities of visiting all paths of length $(b+1)$ from y to k. Therefore, we have that $\mathbb{M}[\Phi^{0,b+1}_{k,y}]=[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b}\mathbf{H}^\mathsf{T}]_{k,y}$. This completes the proof for the third case.

In the fourth case, we have that $0<a<b$. We proceed by induction on both a and b. If $a=1$ and $b=2$, we have that

$$\begin{aligned}&[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a-1}\mathbf{H}^\mathsf{T}]_{x,y}\\&=[\mathbf{H}\mathbf{E}\mathbf{E}^\mathsf{T}\mathbf{H}^\mathsf{T}]_{x,y}=[\mathbf{P}\mathbf{P}^\mathsf{T}]_{x,y}=\textstyle\sum_{i\in V}p_{x,i}\cdot p_{y,i}\end{aligned}$$

Each term $(p_{x,i}\cdot p_{y,i})$ represents the probability of visiting the meeting path $x\to i\leftarrow y$. Thus, $\sum_{i\in V}p_{x,i}\cdot p_{y,i}$ represents the sum of probabilities of visiting the paths in $\Phi^{1,2}_{x,y}$. Therefore, we have that $\mathbb{M}[\Phi^{1,2}_{x,y}]=[\mathbf{H}\mathbf{E}\mathbf{E}^\mathsf{T}\mathbf{H}^\mathsf{T}]_{x,y}$ thus the lemma holds for $\{a=1,b=2\}$. Now assume that the lemma holds for $\{a,b\}\,(0<a<b)$. We will prove that the lemma holds for both $\{a+1,b+1\}$ and $\{a,b+1\}$. By the assumption, we have

$$\begin{aligned}\mathbb{M}[\Phi^{a,b}_{x,y}]&=[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a-1}\mathbf{H}^\mathsf{T}]_{x,y}\\&=\textstyle\sum_{j\in V}[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}]_{x,j}\cdot[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a-1}\mathbf{H}^\mathsf{T}]_{j,y}\end{aligned}$$

We first prove that the lemma holds for $\{a+1,b+1\}$. As discussed in the second case, each term $[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}]_{x,j}=\mathbb{M}[\Phi^{a,a}_{x,j}]$ represents the sum of probabilities of visiting the paths of length a from node x to j. Following the same discussion in the second case, we can prove that $[\mathbf{H}\mathbf{M}^{a}\mathbf{E}]_{x,j}=\mathbb{M}[\Phi^{a+1,a+1}_{x,j}]$ represents the sum of probabilities of visiting the paths of length $(a+1)$ from node x to j. Thus, we have that

$$\begin{aligned}&\textstyle\sum_{j\in V}[\mathbf{H}\mathbf{M}^{a}\mathbf{E}]_{x,j}\cdot[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a-1}\mathbf{H}^\mathsf{T}]_{j,y}\\&=[\mathbf{H}\mathbf{M}^{a}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a-1}\mathbf{H}^\mathsf{T}]_{x,y}=\mathbb{M}[\Phi^{a+1,b+1}_{x,y}]\end{aligned}$$

represents the sum of probabilities of visiting the meeting paths of length $\{a+1,b+1\}$ between nodes x and y. Thus, the lemma holds for $\{a+1,b+1\}$.

We then prove that the lemma holds for $\{a,b+1\}$. As discussed in the third case, each term $[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a-1}\mathbf{H}^\mathsf{T}]_{j,y}=\mathbb{M}[\Phi^{0,b-a}_{j,y}]$ represents the sum of probabilities of visiting the paths of length $(b-a)$ from node y to j. Following the same discussion in the third case, we can prove that $[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a}\mathbf{H}^\mathsf{T}]_{j,y}=\mathbb{M}[\Phi^{0,b-a+1}_{j,y}]$ represents the sum of probabilities of visiting the paths of length $(b-a+1)$ from node y to j. Thus, we have that

$$\begin{aligned}&\textstyle\sum_{j\in V}[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}]_{x,j}\cdot[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a}\mathbf{H}^\mathsf{T}]_{j,y}\\&=[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a}\mathbf{H}^\mathsf{T}]_{x,y}=\mathbb{M}[\Phi^{a,b+1}_{x,y}]\end{aligned}$$

represents the sum of probabilities of visiting the meeting paths of length $\{a,b+1\}$ between nodes x and y. Thus, the lemma holds for $\{a,b+1\}$. This completes the proof for the fourth case. $\square$

The proof of Lemma 3 is as follows.

Proof. If $0=a=b$, the lemma trivially holds, i.e., $\mathbb{M}[\Phi^{0,0}_{i,j}]=\mathbb{P}[\Phi^{0,0}_{i,j}]=\mathbf{I}_{i,j}$. By Lemma 4, if $p_{i,j,k}=p_{j,k}$, we have that $\mathbf{M}=\mathbf{E}\mathbf{H}$. If $0<a=b$, we have that

$$\begin{aligned}\mathbb{M}[\Phi^{a,a}_{i,j}]&=[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}]_{i,j}=[\mathbf{H}(\mathbf{E}\mathbf{H})^{a-1}\mathbf{E}]_{i,j}\\&=[(\mathbf{H}\mathbf{E})^{a}]_{i,j}=[\mathbf{P}^{a}]_{i,j}=\mathbb{P}[\Phi^{a,a}_{i,j}]\end{aligned}$$

If $0=a<b$, we have that

$$\begin{aligned}\mathbb{M}[\Phi^{0,b}_{i,j}]&=[\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{i,j}=[\mathbf{E}^\mathsf{T}(\mathbf{H}^\mathsf{T}\mathbf{E}^\mathsf{T})^{b-1}\mathbf{H}^\mathsf{T}]_{i,j}\\&=[(\mathbf{E}^\mathsf{T}\mathbf{H}^\mathsf{T})^{b}]_{i,j}=[(\mathbf{P}^\mathsf{T})^{b}]_{i,j}=\mathbb{P}[\Phi^{0,b}_{i,j}]\end{aligned}$$

If $0<a<b$, we have that

$$\begin{aligned}\mathbb{M}[\Phi^{a,b}_{i,j}]&=[\mathbf{H}\mathbf{M}^{a-1}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{b-a-1}\mathbf{H}^\mathsf{T}]_{i,j}\\&=[\mathbf{H}(\mathbf{E}\mathbf{H})^{a-1}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{H}^\mathsf{T}\mathbf{E}^\mathsf{T})^{b-a-1}\mathbf{H}^\mathsf{T}]_{i,j}\\&=[(\mathbf{H}\mathbf{E})^{a}(\mathbf{E}^\mathsf{T}\mathbf{H}^\mathsf{T})^{b-a}]_{i,j}=[\mathbf{P}^{a}(\mathbf{P}^\mathsf{T})^{b-a}]_{i,j}=\mathbb{P}[\Phi^{a,b}_{i,j}]\end{aligned}$$

This completes the proof. $\square$

C. THE SECOND-ORDER SIMRANK

The proof of Theorem 2 is as follows.

Proof. The second-order SimRank is defined as

$$r_{i,j}=(1-c)\textstyle\sum_{t=0}^{\infty}c^{t}\mathbb{M}[\Phi^{t,2t}_{i,j}]$$

By Lemma 2, we have that $\mathbb{M}[\Phi^{0,0}_{i,j}]=\mathbf{I}_{i,j}$ and $\mathbb{M}[\Phi^{t,2t}_{i,j}]=[\mathbf{H}\mathbf{M}^{t-1}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{t-1}\mathbf{H}^\mathsf{T}]_{i,j}$ if $t>0$. Thus, the node proximity matrix $\mathbf{R}$ can be expressed as

$$\begin{aligned}\mathbf{R}&=(1-c)\textstyle\sum_{t=1}^{\infty}c^{t}\mathbf{H}\mathbf{M}^{t-1}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{t-1}\mathbf{H}^\mathsf{T}+(1-c)\mathbf{I}\\&=c\mathbf{H}\big((1-c)\textstyle\sum_{t=1}^{\infty}c^{t-1}\mathbf{M}^{t-1}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{t-1}\big)\mathbf{H}^\mathsf{T}+(1-c)\mathbf{I}\\&=c\mathbf{H}\big((1-c)\textstyle\sum_{t=0}^{\infty}c^{t}\mathbf{M}^{t}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{t}\big)\mathbf{H}^\mathsf{T}+(1-c)\mathbf{I}\end{aligned}$$

Let $\mathbf{S}=(1-c)\sum_{t=0}^{\infty}c^{t}\mathbf{M}^{t}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{t}$. Thus, we have that $\mathbf{R}=c\mathbf{H}\mathbf{S}\mathbf{H}^\mathsf{T}+(1-c)\mathbf{I}$. Matrix $\mathbf{S}$ can be written as

$$\begin{aligned}&c\mathbf{M}\mathbf{S}\mathbf{M}^\mathsf{T}+(1-c)\mathbf{E}\mathbf{E}^\mathsf{T}\\&=(1-c)\textstyle\sum_{t=0}^{\infty}c^{t+1}\mathbf{M}^{t+1}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{t+1}+(1-c)\mathbf{E}\mathbf{E}^\mathsf{T}\\&=(1-c)\textstyle\sum_{t=1}^{\infty}c^{t}\mathbf{M}^{t}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{t}+(1-c)\mathbf{E}\mathbf{E}^\mathsf{T}\\&=(1-c)\textstyle\sum_{t=0}^{\infty}c^{t}\mathbf{M}^{t}\mathbf{E}\mathbf{E}^\mathsf{T}(\mathbf{M}^\mathsf{T})^{t}=\mathbf{S}\end{aligned}\qquad\square$$

Lemma 5 is needed in the proof of Theorem 3.

Lemma 5. *The gap between* $\mathbf{S}$ *and* $\mathbf{S}^{(\eta)}$ *is bounded by* $\left\|\mathbf{S}-\mathbf{S}^{(\eta)}\right\|_{\max}\leq c^{\eta+1}$ *for any* $\eta\,(\eta\geq 0)$.

Proof. For each $\eta=0,1,\cdots$, we subtract $\mathbf{S}^{(\eta)}$ from $\mathbf{S}$, and then take $\|\cdot\|_{\max}$ norms on both sides to get

Table 5: The sample space when the length a is given

sample space		$\mathbb{S}_i$
bSuccess	node z_a	
successfully sample a path of length a starting from node q (bSuccess = true)	$z_a = i$	1
	$z_a \neq i$	0
fail to sample a path of length a starting from node q (bSuccess = false)	–	0

$$\left\|\mathbf{S}-\mathbf{S}^{(\eta)}\right\|_{\max} \le (1-c)\sum_{t=\eta+1}^{\infty} c^t \left\|\mathbf{M}^t\mathbf{E}\mathbf{E}^{\intercal}(\mathbf{M}^{\intercal})^t\right\|_{\max}$$

Note that matrix $\mathbf{E}\mathbf{E}^{\intercal}$ is binary. Each element $[\mathbf{E}\mathbf{E}^{\intercal}]_{u,v}=1$ if edge u and v end at the same node and $[\mathbf{E}\mathbf{E}^{\intercal}]_{u,v}=0$ otherwise. Thus, we have that $\left\|\mathbf{M}^t\mathbf{E}\mathbf{E}^{\intercal}(\mathbf{M}^{\intercal})^t\right\|_{\max}\le 1$. Plugging this into the above inequality, we have that

$$\left\|\mathbf{S}-\mathbf{S}^{(\eta)}\right\|_{\max} \le (1-c)\sum_{t=\eta+1}^{\infty} c^t = c^{\eta+1} \qquad \square$$

D. MONTE CARLO METHODS

Lemma 6 is needed in the proof of Theorem 4.

LEMMA 6. $\mathbb{E}[\mathbb{S}_i] = r_i$

PROOF. The set of all possible outcomes of the sampling process is called the sample space, which contains the following events.

1) The algorithm generates a random number a and successfully samples a path of length a starting from the query node q.
2) The algorithm generates a random number a but fails to sample a path of length a starting from the query node q because some node has no out-neighbors.

Note that this sample space is different from the sample space of the random variable $\mathbb{S}_i$, which is the set of two integers $\{0,1\}$.

The whole sample space can be partitioned based on the length a. By the law of total expectation, the expectation of $\mathbb{S}_i$ can be written as

$$\mathbb{E}[\mathbb{S}_i] = \sum_{a=0}^{\infty} \mathbb{P}[\mathbb{A}=a]\cdot\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a], \tag{2}$$

where the random variable $\mathbb{A}$ representing the length a follows the geometric distribution $\mathbb{P}[\mathbb{A}=a]=(1-c)\cdot c^a$. Next, we consider the conditional expectation $\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a]$ of $\mathbb{S}_i$ given the event $\mathbb{A}=a$.

Table 5 shows the sample space given the length a. Given the length a, if the algorithm successfully samples a path of length a from node q to i, the random variable $\mathbb{S}_i=1$; otherwise, $\mathbb{S}_i=0$. Let $\underline{\mathbb{M}}[\rho]$ represent the probability of successfully sampling a path ρ given the length a. Since $\mathbb{S}_i=1$ if and only if the algorithm successfully samples a path ρ of length a from node q to i, the conditional expectation $\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a]$ can be written as

$$\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a] = \sum_{\rho\in\Phi_{q,i}^{a,a}} 1\cdot\underline{\mathbb{M}}[\rho] = \sum_{\rho\in\Phi_{q,i}^{a,a}} \underline{\mathbb{M}}[\rho],$$

where $\Phi_{q,i}^{a,a}$ denotes the set of all paths of length a from node q to i.

Given the length a, the probability of successfully sampling a path $\rho: q=z_0\to\cdots\to z_a$ is $\underline{\mathbb{M}}[\rho]=p_{z_0,z_1}\prod_{t=1}^{a-1}p_{z_{t-1},z_t,z_{t+1}}$. We can see that the probabilities of sampling and visiting a path ρ are equal, i.e., $\underline{\mathbb{M}}[\rho]=\mathbb{M}[\rho]$. Thus, we have that

$$\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a] = \sum_{\rho\in\Phi_{q,i}^{a,a}} \mathbb{M}[\rho] = \mathbb{M}[\Phi_{q,i}^{a,a}]$$

Table 6: The sample space when the length a is given

a	sample space		$\mathbb{R}_i$
	bSuccess	node z_{2a}	
$0\le a\le\eta$	successfully sample a path of length a starting from node q (bSuccess = true)	$z_{2a}=i$	1
		$z_{2a}\neq i$	0
	fail to sample a path of length a starting from node q (bSuccess = false)	–	0
$\eta<a$	–	–	0

Plugging this into Equation (2), we have that

$$\mathbb{E}[\mathbb{S}_i] = (1-c)\sum_{a=0}^{\infty} c^a\mathbb{M}[\Phi_{q,i}^{a,a}] = r_i \qquad \square$$

Lemma 7 and Theorem 14 are needed in the proofs of Theorems 7 and 8.

LEMMA 7. $\mathbb{E}[\mathbb{R}_i]=\hat{r}_i$ *and* $\mathbb{E}[\mathbb{R}_i^2]\le nr_i$

PROOF. The set of all possible outcomes of the sampling process is called the sample space, which contains the following events.

1) The algorithm generates a random number $a\,(0\le a\le\eta)$ and successfully samples a meeting path of length $\{a,2a\}$ starting from the query node q.
2) The algorithm generates a random number $a\,(0\le a\le\eta)$ but fails to sample a meeting path of length $\{a,2a\}$ starting from the query node q because some node has no out-neighbors or in-neighbors.
3) The algorithm generates a random number $a\,(\eta<a)$ and does nothing.

Note that this sample space is different from the sample space of the random variable $\mathbb{R}_i$, which is the set of real values $\{\delta\}$.

The whole sample space can be partitioned based on the length a. By the law of total expectation, the expectation of $\mathbb{R}_i$ can be written as

$$\mathbb{E}[\mathbb{R}_i] = \sum_{a=0}^{\infty} \mathbb{P}[\mathbb{A}=a]\cdot\mathbb{E}[\mathbb{R}_i|\mathbb{A}=a], \tag{3}$$

where the random variable $\mathbb{A}$ representing the length a follows the geometric distribution $\mathbb{P}[\mathbb{A}=a]=(1-c)\cdot c^a$. Next, we consider the conditional expectation $\mathbb{E}[\mathbb{R}_i|\mathbb{A}=a]$ of $\mathbb{R}_i$ given the event $\mathbb{A}=a$.

Table 6 shows the sample space given the length a. Given the length $a\,(0\le a\le\eta)$, if the algorithm successfully samples a meeting path of length $\{a,2a\}$ between nodes q and i, the random variable $\mathbb{R}_i=\delta$; otherwise, $\mathbb{R}_i=0$. Note that $\delta=[\mathbf{X}]_{z_a,a}/[\mathbf{X}]_{z_{2a},0}$ changes for different sampled meeting paths. Let $\underline{\mathbb{P}}[\phi]$ represent the probability of successfully sampling a meeting path ϕ given the length a. Since $\mathbb{R}_i=\delta$ if and only if the algorithm successfully samples a meeting path ϕ of length $\{a,2a\}$ between nodes q and i, the conditional expectation $\mathbb{E}[\mathbb{R}_i|\mathbb{A}=a]$ can be written as

$$\mathbb{E}[\mathbb{R}_i|\mathbb{A}=a] = \sum_{\phi\in\Phi_{q,i}^{a,2a}} \delta\cdot\underline{\mathbb{P}}[\phi],$$

where $\Phi_{q,i}^{a,2a}$ denotes the set of all meeting paths of length $\{a,2a\}$ between nodes q and i.

Consider the probability of successfully sampling a meeting path $\phi: q=z_0\to\cdots\to z_a\leftarrow\cdots\leftarrow z_{2a}$ given the length a. The probability of sampling the first half is $\underline{\mathbb{P}}[\rho_1]=\prod_{t=1}^{a}p_{z_{t-1},z_t}$. The probability of sampling the second half is

$$\underline{\mathbb{P}}[\rho_2] = \prod_{t=a+1}^{2a}\left(p_{z_t,z_{t-1}}\cdot\frac{[\mathbf{X}]_{z_t,2a-t}}{[\mathbf{X}]_{z_{t-1},2a-t+1}}\right) = \frac{[\mathbf{X}]_{z_{2a},0}}{[\mathbf{X}]_{z_a,a}}\cdot\prod_{t=a+1}^{2a}p_{z_t,z_{t-1}}$$

The probability of sampling the meeting path ϕ then is $\underline{\mathbb{P}}[\phi] = \underline{\mathbb{P}}[\rho_1] \cdot \underline{\mathbb{P}}[\rho_2]$. Note that $\delta = [\mathbf{X}]_{z_a,a} / [\mathbf{X}]_{z_{2a},0}$. We can see that the probabilities of sampling and visiting a meeting path ϕ have a relationship, i.e., $\delta \cdot \underline{\mathbb{P}}[\phi] = \mathbb{P}[\phi]$. Thus, if $0 \le a \le \eta$, we have

$$\mathbb{E}[\mathbb{R}_i | \mathbb{A} = a] = \sum_{\phi \in \Phi_{q,i}^{a,2a}} \mathbb{P}[\phi] = \mathbb{P}[\Phi_{q,i}^{a,2a}]$$

If $\eta < a$, we have $\mathbb{E}[\mathbb{R}_i | \mathbb{A} = a] = 0$. Plugging this into Equation (3), we have that

$$\mathbb{E}[\mathbb{R}_i] = (1-c) \sum_{a=0}^{\eta} c^a \mathbb{P}[\Phi_{q,i}^{a,2a}] = \hat{r}_i \ ,$$

where $\hat{r}_i$ is the truncated SimRank proximity.

Next, we prove that $\mathbb{E}[\mathbb{R}_i^2] \le n r_i$. By the law of total expectation, we have that

$$\mathbb{E}[\mathbb{R}_i^2] = \sum_{a=0}^{\infty} \mathbb{P}[\mathbb{A} = a] \cdot \mathbb{E}[\mathbb{R}_i^2 | \mathbb{A} = a] \tag{4}$$

Since $[\mathbf{X}]_{z_a,a} \in [0,1]$ and $[\mathbf{X}]_{z_{2a},0} = \frac{1}{n}$, where n is the number of nodes in the graph, we have that $\delta = [\mathbf{X}]_{z_a,a} / [\mathbf{X}]_{z_{2a},0} \le n$. Thus, if $0 \le a \le \eta$, we have

$$\mathbb{E}[\mathbb{R}_i^2 | \mathbb{A} = a] = \sum_{\phi \in \Phi_{q,i}^{a,2a}} \delta^2 \, \underline{\mathbb{P}}[\phi] = \sum_{\phi \in \Phi_{q,i}^{a,2a}} \delta \, \mathbb{P}[\phi] \le n \mathbb{P}[\Phi_{q,i}^{a,2a}]$$

If $\eta < a$, we have $\mathbb{E}[\mathbb{R}_i^2 | \mathbb{A} = a] = 0$. Plugging this into Equation (4), we have that

$$\mathbb{E}[\mathbb{R}_i^2] \le (1-c) \sum_{a=0}^{\infty} c^a n \mathbb{P}[\Phi_{q,i}^{a,2a}] = n r_i \qquad \square$$

Theorem 14 is based on Theorems 2.8 and 2.9 in [5].

THEOREM 14. [Concentration Inequality] *Let* $\mathbb{U}_1, \cdots, \mathbb{U}_\pi$ *be independent random variables bounded by interval* $[-\alpha, \beta]$, *where* α *and* β *are non-negative constants, i.e.,* $-\alpha \le \mathbb{U}_d \le \beta$ *for each* $d = 1, \cdots, \pi$. *Let* $\mathbb{V} = \frac{1}{\pi} \sum_{d=1}^{\pi} \mathbb{U}_d$ *and* $\theta = \sum_{d=1}^{\pi} \mathbb{E}[\mathbb{U}_d^2]$. *For any* $\epsilon > 0$, *we have that*

$$\begin{cases} \mathbb{P}[\mathbb{V} - \mathbb{E}[\mathbb{V}] \le -\epsilon] \le \exp\left(\dfrac{-\pi^2 \epsilon^2}{2\theta + 2\pi\alpha\epsilon/3} \right) \\ \mathbb{P}[\mathbb{V} - \mathbb{E}[\mathbb{V}] \ge \ \epsilon \ ] \le \exp\left(\dfrac{-\pi^2 \epsilon^2}{2\theta + 2\pi\beta\epsilon/3} \right) \end{cases}$$