Appendix

This is the appendix for the paper titled "Second-order random walk based proximity measures in graph analysis: Formulations and algorithms".

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 $\mathbf{M}=\mathbf{EH}$.*

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 $\mathbf{M}=\mathbf{EH}$. □

Theorem 17 *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

$$\mathbf{r}=\mathbf{P}^\top\mathbf{r}=\mathbf{E}^\top\mathbf{H}^\top\mathbf{r}$$

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

$$\mathbf{H}^\top\mathbf{r}=\mathbf{H}^\top\mathbf{E}^\top\mathbf{H}^\top\mathbf{r}$$

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

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

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 18 *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

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

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

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

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

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

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

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 19 *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 In the first-order random walk with restart, the recursive equation is

$$\mathbf{r}=c\mathbf{P}^\top\mathbf{r}+(1-c)\mathbf{q}=c\mathbf{E}^\top\mathbf{H}^\top\mathbf{r}+(1-c)\mathbf{q}$$

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

$$\mathbf{H}^\top\mathbf{r}=c\mathbf{H}^\top\mathbf{E}^\top\mathbf{H}^\top\mathbf{r}+(1-c)\mathbf{H}^\top\mathbf{q}$$

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

$$\mathbf{s}=c\mathbf{H}^\top\mathbf{E}^\top\mathbf{s}+(1-c)\mathbf{H}^\top\mathbf{q}=c\mathbf{M}^\top\mathbf{s}+(1-c)\mathbf{H}^\top\mathbf{q}$$

and $$\mathbf{r}=c\mathbf{E}^\top\mathbf{H}^\top\mathbf{r}+(1-c)\mathbf{q}=c\mathbf{E}^\top\mathbf{s}+(1-c)\mathbf{q}$$

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

Theorem 20 *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 21 *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 The second-order SimRank

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

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

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

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

Each term $[(\mathbf{M}^\intercal)^{b-1}\mathbf{H}^\intercal]_{(i,j),y}$ 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 $[\mathbf{E}^\intercal(\mathbf{M}^\intercal)^{b}\mathbf{H}^\intercal]_{k,y}$ can be expanded as

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

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 $p_{i,j,k}\cdot[(\mathbf{M}^\intercal)^{b-1}\mathbf{H}^\intercal]_{(i,j),y}$ 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, $\sum_{j\in I_k}\sum_{i\in I_j}p_{i,j,k}\cdot[(\mathbf{M}^\intercal)^{b-1}\mathbf{H}^\intercal]_{(i,j),y}$ is the sum of probabilities of visiting all paths of length $(b+1)$ from y to k. Therefore, we have that $\mathbb{M}[\Phi_{k,y}^{0,b+1}]=[\mathbf{E}^\intercal(\mathbf{M}^\intercal)^{b}\mathbf{H}^\intercal]_{k,y}$. 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

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

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 $\Phi_{x,y}^{1,2}$. Therefore, we have that $\mathbb{M}[\Phi_{x,y}^{1,2}]=[\mathbf{H}\mathbf{E}\mathbf{E}^\intercal\mathbf{H}^\intercal]_{x,y}$ 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

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

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

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

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 $[\mathbf{E}^{\mathsf{T}}(\mathbf{M}^{\mathsf{T}})^{b-a-1}\mathbf{H}^{\mathsf{T}}]_{j,y} = \mathbb{M}[\Phi_{j,y}^{0,b-a}]$ 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 $[\mathbf{E}^{\mathsf{T}}(\mathbf{M}^{\mathsf{T}})^{b-a}\mathbf{H}^{\mathsf{T}}]_{j,y} = \mathbb{M}[\Phi_{j,y}^{0,b-a+1}]$ represents the sum of probabilities of visiting the paths of length $(b-a+1)$ from node y to j. Thus, we have that

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

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. □

The proof of Lemma 3 is as follows.

Proof If $0=a=b$, the lemma trivially holds, i.e., $\mathbb{M}[\Phi_{i,j}^{0,0}] = \mathbb{P}[\Phi_{i,j}^{0,0}] = \mathbf{I}_{i,j}$. 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

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

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

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

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

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

This completes the proof. □

The proof of Theorem 2 is as follows.

Proof The second-order SimRank is defined as

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

By Lemma 2, we have that $\mathbb{M}[\Phi_{i,j}^{0,0}] = \mathbf{I}_{i,j}$ and $\mathbb{M}[\Phi_{i,j}^{t,2t}] = [\mathbf{H}\mathbf{M}^{t-1}\mathbf{E}\mathbf{E}^{\mathsf{T}}(\mathbf{M}^{\mathsf{T}})^{t-1}\mathbf{H}^{\mathsf{T}}]_{i,j}$ 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

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

This completes the proof of Theorem 2. □

Lemma 5 *The gap between* $\mathbf{S}$ *and* $\mathbf{S}^{(\eta)}$ *is bounded by* $\big\|\mathbf{S}-\mathbf{S}^{(\eta)}\big\|_{\max} \le c^{\eta+1}$ *for any* η $(\eta \ge 0)$.

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

$$\big\|\mathbf{S}-\mathbf{S}^{(\eta)}\big\|_{\max} \le (1-c)\textstyle\sum_{t=\eta+1}^{\infty} c^{t}\big\|\mathbf{M}^{t}\mathbf{E}\mathbf{E}^{\mathsf{T}}(\mathbf{M}^{\mathsf{T}})^{t}\big\|_{\max}$$

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

$$\big\|\mathbf{S}-\mathbf{S}^{(\eta)}\big\|_{\max} \le (1-c)\textstyle\sum_{t=\eta+1}^{\infty} c^{t} = c^{\eta+1}$$

This completes the proof of Lemma 5. □

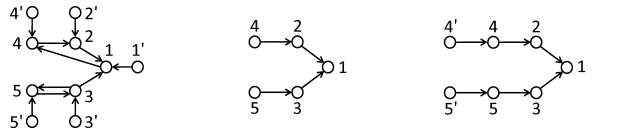
C The second-order SimRank*

The proof of Theorem 4 is as follows.

Proof To transform Equation (4) to the matrix form, we represent Lemma 2 in another way by using a leaf-augmented graph.

Definition 2 *[Leaf-Augmented Graph]* Given a graph $G(V,E)$, its leaf-augmented graph $G'(V',E')$ is constructed as follows.

1) for each node $i \in V$, add a node i to V';
2) for each edge $(i,j) \in E$, add an edge (i,j) to E';
3) for each node $i \in V$, add a node i' to V' and an edge (i',i) to E'.



(a) leaf-augmented graph (b) a meeting path in the original graph (c) leaf-augmented meeting path

Fig. 23 Leaf-augmented graph and meeting path

For each node i in the original graph, we add a "shadow" node i' in the leaf-augmented graph. Node i' is a leaf node since it is only connected to node i. Figure 23(a) shows the leaf-augmented graph of the graph in Figure 4(a).

Definition 3 *[Leaf-Augmented Meeting Path]* Given a meeting path $\phi: i=z_0 \to z_1 \to \cdots \to z_a \leftarrow \cdots \leftarrow z_{b-1} \leftarrow z_b = j$ of length $\{a,b\}$ between nodes i and j in graph $G(V,E)$, the leaf-augmented meeting path of ϕ is defined as the meeting path $\phi': i'=z_{-1} \to z_0 \to z_1 \to \cdots \to z_a \leftarrow \cdots \leftarrow z_{b-1} \leftarrow z_b \leftarrow z_{b+1}=j'$ of length $\{a+1, b+2\}$ between nodes i' and j' in the leaf-augmented graph $G'(V',E')$ of $G(V,E)$.

For each meeting path in the original graph $G(V,E)$, there is one and only one leaf-augmented meeting path in the leaf-augmented graph $G'(V',E')$ of $G(V,E)$. For example, Figure 23(b) shows a meeting path in the graph in Figure 4(a) and Figure 23(c) shows its leaf-augmented meeting path.

In the leaf-augmented graph $G'(V',E')$, we keep the first-order transition probability $p_{i,j}$ and the second-order transition probability $p_{i,j,k}$ among the nodes in the original graph $G(V,E)$. For the newly added nodes and edges, the transition probabilities are defined as follows.

1) For any node $i' \in V'$, set the first-order transition probability $p_{i',i}=1$;
2) For any node $i' \in V'$ and $j \in O_i$, set the second-order transition probability $p_{i',i,j}=p_{i,j}$.

Recall that we use matrix $\mathbf{H}$ to denote the node-to-edge transition probability matrix, and matrix $\mathbf{M}$ to denote the edge-to-edge transition probability matrix in the original graph $G(V,E)$. For the leaf-augmented graph $G'(V',E')$, we use matrix $\mathbf{H}'$ to denote the node-to-edge transition probability matrix, and matrix $\mathbf{M}'$ to denote the edge-to-edge transition probability matrix. Matrices $\mathbf{H}'$ and $\mathbf{M}'$ can be represented by 2×2 block matrices as $\mathbf{H}'=\begin{bmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}$ and $\mathbf{M}'=\begin{bmatrix}\mathbf{M} & \mathbf{0}\\ \mathbf{H} & \mathbf{0}\end{bmatrix}$, where $\mathbf{0}$ is a matrix of all 0's. The partition of blocks for matrix $\mathbf{H}'$ is $(n|n)\times(m|n)$ and that for matrix $\mathbf{M}'$ is $(m|n)\times(m|n)$. We also use $\mathbf{E}'=\begin{bmatrix}\mathbf{E} & \mathbf{0}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}$ to denote the in-edge incidence matrix of the leaf-augmented graph and the partition of blocks is $(m|n)\times(n|n)$.

In the leaf-augmented meeting path $\phi': i'=z_{-1} \to z_0 \to z_1 \to \cdots \to z_a \leftarrow \cdots \leftarrow z_{b-1} \leftarrow z_b \leftarrow z_{b+1}=j'$, we can see that the transition probability from z_{-1} to z_0 and that from z_{b+1} to z_b are both 1, i.e., $p_{z_{-1},z_0}=p_{z_{b+1},z_b}=1$. We also have that $p_{z_{-1},z_0,z_1}=p_{z_0,z_1}$ and $p_{z_{b+1},z_b,z_{b-1}}=p_{z_b,z_{b-1}}$. Thus, the probability to visit the leaf-augmented meeting path ϕ' is equal to that to visit the original meeting path ϕ.

Based on this observation, Lemma 2 can be represented in another form in Lemma 6.

Lemma 6 $\mathbb{M}[\Phi_{i,j}^{a,b}]=[(\mathbf{M}')^a\mathbf{E}'(\mathbf{E}')^\top((\mathbf{M}')^\top)^{b-a}]_{(i',i),(j',j)}$

Proof Since $\mathbf{M}'$ can be represented as a 2×2 block matrix, we have the following equations.

$$(\mathbf{M}')^a=\begin{cases}\mathbf{I}, & \text{if } 0=a,\\ \begin{bmatrix}\mathbf{M}^a & \mathbf{0}\\ \mathbf{H}\mathbf{M}^{a-1} & \mathbf{0}\end{bmatrix}, & \text{if } 0<a.\end{cases}$$

$$((\mathbf{M}')^\top)^{b-a}=\begin{cases}\mathbf{I}, & \text{if } a=b,\\ \begin{bmatrix}(\mathbf{M}^\top)^{b-a} & (\mathbf{M}^\top)^{b-a-1}\mathbf{H}^\top\\ \mathbf{0} & \mathbf{0}\end{bmatrix}, & \text{if } a<b.\end{cases}$$

Let $\mathbf{Z}=(\mathbf{M}')^a\mathbf{E}'(\mathbf{E}')^\top((\mathbf{M}')^\top)^{b-a}$. Matrix $\mathbf{Z}$ can be represented as 2×2 block matrices in four different cases. If $0=a=b$, we have

$$\mathbf{Z}=\mathbf{E}'(\mathbf{E}')^\top=\begin{bmatrix}\mathbf{E}\mathbf{E}^\top & \mathbf{E}\\ \mathbf{E}^\top & \mathbf{I}\end{bmatrix}$$

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

$$\mathbf{Z}=(\mathbf{M}')^a\mathbf{E}'(\mathbf{E}')^\top=\begin{bmatrix}\mathbf{M}^a\mathbf{E}\mathbf{E}^\top & \mathbf{M}^a\mathbf{E}\\ \mathbf{H}\mathbf{M}^{a-1}\mathbf{E}\mathbf{E}^\top & \mathbf{H}\mathbf{M}^{a-1}\mathbf{E}\end{bmatrix}$$

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

$$\mathbf{Z}=\mathbf{E}'(\mathbf{E}')^\top((\mathbf{M}')^\top)^b=\begin{bmatrix}\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^b & \mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{b-1}\mathbf{H}^\top\\ \mathbf{E}^\top(\mathbf{M}^\top)^b & \mathbf{E}^\top(\mathbf{M}^\top)^{b-1}\mathbf{H}^\top\end{bmatrix}$$

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

$$\mathbf{Z}=\begin{bmatrix}\mathbf{M}^a\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{b-a} & \mathbf{M}^a\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{b-a-1}\mathbf{H}^\top\\ \mathbf{H}\mathbf{M}^{a-1}\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{b-a} & \mathbf{H}\mathbf{M}^{a-1}\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{b-a-1}\mathbf{H}^\top\end{bmatrix}$$

In each of the above four cases, the matrix in the bottom right block exactly matches that in the right-hand side of the equation in Lemma 2. This completes the proof of Lemma 6. □

Plugging Lemma 6 into Equation (4), we can have the following matrix form.

$$\mathbf{S}'=(1-c)\sum_{t=0}^{\infty}\frac{c^t}{2^t}\sum_{a=0}^{t}\binom{t}{a}(\mathbf{M}')^a\mathbf{E}'(\mathbf{E}')^\top((\mathbf{M}')^\top)^{t-a} \quad (12)$$

Matrix $\mathbf{S}'$ is symmetric and can be denoted by a 2×2 block matrix as $\mathbf{S}'=\begin{bmatrix}\mathbf{S} & \mathbf{R}'\\ (\mathbf{R}')^\top\mathbf{R} & \end{bmatrix}$, where the partition is $(m|n)\times(m|n)$. The matrix $\mathbf{R}$ in the bottom right block denotes the node proximity matrix with each element

$\mathbf{R}_{i,j} = r_{i,j}$. Based on this series form, we can derive a recursive form for $\mathbf{S}'$ as

$$\mathbf{S}' = c(\mathbf{M}'\mathbf{S}' + \mathbf{S}'(\mathbf{M}')^\top)/2 + (1-c)\mathbf{E}'(\mathbf{E}')^\top \quad (13)$$

By recursively substituting the left hand side of Equation (13) into the right hand side, we can obtain the series form in Equation (12). This proves the correctness of Equation (13). Plugging the 2×2 block matrix representations into Equation (13), we have

$$\begin{bmatrix} \mathbf{S} & \mathbf{R}' \\ (\mathbf{R}')^\top \mathbf{R} \end{bmatrix} = \frac{c}{2}\begin{bmatrix} \mathbf{MS} & \mathbf{MR}' \\ \mathbf{HS} & \mathbf{HR}' \end{bmatrix} + \frac{c}{2}\begin{bmatrix} \mathbf{SM}^\top & \mathbf{SH}^\top \\ (\mathbf{R}')^\top \mathbf{M}^\top & (\mathbf{R}')^\top \mathbf{H}^\top \end{bmatrix} + (1-c)\begin{bmatrix} \mathbf{EE}^\top \mathbf{E} \\ \mathbf{E}^\top & \mathbf{I} \end{bmatrix}$$

By considering each block individually, we can obtain the set of three equations in Theorem 4. This completes the proof of Theorem 4. □

The proof of Theorem 5 is as follows.

Proof By recursively substituting the left hand side of Equation (5) into the right hand side, we obtain the series expansion:

$$\mathbf{S} = (1-c)\sum_{t=0}^{\infty} \frac{c^t}{2^t} \sum_{a=0}^{t} \binom{t}{a} \mathbf{M}^a \mathbf{EE}^\top (\mathbf{M}^\top)^{t-a}$$

Since this is an infinite series, it is unclear whether this series is convergent. Let us define the partial sum as

$$\mathbf{S}^{(\eta)} = (1-c)\sum_{t=0}^{\eta} \frac{c^t}{2^t} \sum_{a=0}^{t} \binom{t}{a} \mathbf{M}^a \mathbf{EE}^\top (\mathbf{M}^\top)^{t-a}$$

Lemma 7 *The gap between* $\mathbf{S}$ *and* $\mathbf{S}^{(\eta)}$ *is bounded by* $\left\|\mathbf{S} - \mathbf{S}^{(\eta)}\right\|_{\max} \le c^{\eta+1}$ *for any* η $(0 \le \eta)$.

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

$$\left\|\mathbf{S} - \mathbf{S}^{(\eta)}\right\|_{\max} \le (1-c)\sum_{t=\eta+1}^{\infty} \frac{c^t}{2^t} \sum_{a=0}^{t} \binom{t}{a} \left\|\mathbf{M}^a \mathbf{EE}^\top (\mathbf{M}^\top)^{t-a}\right\|_{\max}$$

Since matrix $\mathbf{EE}^\top$ is binary, we have that

$$\left\|\mathbf{M}^a \mathbf{EE}^\top (\mathbf{M}^\top)^{t-a}\right\|_{\max} \le 1$$

Plugging this into the above inequality, we have that

$$\left\|\mathbf{S} - \mathbf{S}^{(\eta)}\right\|_{\max} \le (1-c)\sum_{t=\eta+1}^{\infty} \frac{c^t}{2^t} \sum_{a=0}^{t} \binom{t}{a} = c^{\eta+1}$$

This completes the proof of Lemma 7. □

The convergence of the series follows directly from Lemma 7 and $\lim_{\eta\to\infty} c^{\eta+1} = 0$ $(0 < c < 1)$. Thus, the solution $\mathbf{S}$ exists.

In the following, we show that the solution $\mathbf{R}'$ to Equation (6) also exists. From Equation (6), we have

$$(\mathbf{I} - \tfrac{c}{2}\mathbf{M})\mathbf{R}' = \tfrac{c}{2}\mathbf{SH}^\top + (1-c)\mathbf{E}$$

Since we have that

$$\left\|(\tfrac{c}{2}\mathbf{M})^t\right\|_{\max} \le (\tfrac{c}{2})^t \left\|\mathbf{M}^t\right\|_{\max} \le (\tfrac{c}{2})^t ,$$

Thus, we have

$$\lim_{t\to\infty} \left\|(\tfrac{c}{2}\mathbf{M})^t\right\|_{\max} = 0 \text{ and } \lim_{t\to\infty} (\tfrac{c}{2}\mathbf{M})^t = \mathbf{0},$$

where $\mathbf{0}$ is a matrix of all 0's. Thus matrix $(\mathbf{I} - \frac{c}{2}\mathbf{M})$ is invertible [23]. Given that matrix $\mathbf{S}$ exists, matrix $\mathbf{R}'$ exists.

Next, we prove that the solution $\mathbf{S}$ is unique. Suppose that $\mathbf{S}$ and $\mathbf{S}'$ are two solutions to Equation (5).

$$\begin{cases} \mathbf{S} = c(\mathbf{MS} + \mathbf{SM}^\top)/2 + (1-c)\mathbf{EE}^\top \\ \mathbf{S}' = c(\mathbf{MS}' + \mathbf{S}'\mathbf{M}^\top)/2 + (1-c)\mathbf{EE}^\top \end{cases}$$

Let $\Delta = \mathbf{S} - \mathbf{S}'$ be the difference. We have $\Delta = \frac{c}{2}(\mathbf{M}\Delta + \Delta\mathbf{M}^\top)$. Let $|\Delta_{u,v}| = \|\Delta\|_{\max}$ for some $u, v \in E$.

$$\begin{aligned} &\|\Delta\|_{\max} = |\Delta_{u,v}| = \tfrac{c}{2} \cdot |\mathbf{M}_{u,:} \cdot \Delta_{:,v} + \Delta_{u,:} \cdot (\mathbf{M}_{v,:})^\top| \\ &\le \tfrac{c}{2} \textstyle\sum_{x \in O_u} p_{u,x} \cdot |\Delta_{x,v}| + \tfrac{c}{2} \sum_{y \in O_v} p_{v,y} \cdot |\Delta_{u,y}| \\ &\le \tfrac{c}{2} \big(\textstyle\sum_{x \in O_u} p_{u,x} + \sum_{y \in O_v} p_{v,y} \big) \cdot \|\Delta\|_{\max} = c \cdot \|\Delta\|_{\max} \end{aligned}$$

Since $0 < c < 1$, we have that $\|\Delta\|_{\max} = 0$ and $\mathbf{S} = \mathbf{S}'$. Thus, the solution $\mathbf{S}$ is unique. Since Equation (6) is a system of linear equations, given that $\mathbf{S}$ is unique, $\mathbf{R}'$ is also unique. Since Equation (7) is a system of linear equations, given that $\mathbf{R}'$ exists and is unique, $\mathbf{R}$ also exists and is unique. This completes the proof of Theorem 5. □

D The single-source algorithm

D.1 The first-order SimRank

By recursively substituting the left hand side into the right hand side of the recursive equation of $\mathbf{R}$, we obtain the following series expansion:

$$\mathbf{R} = (1-c)\sum_{t=0}^{\infty} c^t \mathbf{P}^t (\mathbf{P}^\top)^t.$$

The q-th column vector $\mathbf{r} = \mathbf{Rq}$ can be represented as

$$\mathbf{r} = (1-c)\sum_{t=0}^{\infty} c^t \mathbf{P}^t (\mathbf{P}^\top)^t \mathbf{q}.$$

We truncate the series to approximate the original proximity. Let η denote the truncated length and $\hat{\mathbf{r}}$ denote the truncated proximity. We have that

$$\hat{\mathbf{r}} = (1-c)\sum_{t=0}^{\eta} c^t \mathbf{P}^t (\mathbf{P}^\top)^t \mathbf{q}.$$

D.2 The second-order SimRank

By recursively substituting the left hand side into the right hand side of the recursive equation of $\mathbf{S}$, we obtain the following series expansion:

$$\mathbf{S}=(1-c)\sum_{t=0}^{\infty}c^t\mathbf{M}^t\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^t.$$

Plugging this into the equation $\mathbf{R}=c\mathbf{H}\mathbf{S}\mathbf{H}^\top+(1-c)\mathbf{I}$, we have that

$$\mathbf{R}=(1-c)\sum_{t=1}^{\infty}c^t\mathbf{H}\mathbf{M}^{t-1}\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{t-1}\mathbf{H}^\top+(1-c)\mathbf{I}.$$

The q-th column vector $\mathbf{r}=\mathbf{R}\mathbf{q}$ can be represented as

$$\mathbf{r}=(1-c)\sum_{t=1}^{\infty}c^t\mathbf{H}\mathbf{M}^{t-1}\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{t-1}\mathbf{H}^\top\mathbf{q}+(1-c)\mathbf{q}.$$

We truncate the series to approximate the original proximity. Let η denote the truncated length and $\hat{\mathbf{r}}$ denote the truncated proximity. We have that

$$\hat{\mathbf{r}}=(1-c)\sum_{t=1}^{\eta}c^t\mathbf{H}\mathbf{M}^{t-1}\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{t-1}\mathbf{H}^\top\mathbf{q}+(1-c)\mathbf{q}.$$

D.3 The first-order SimRank*

By recursively substituting the left hand side into the right hand side of the recursive equation of $\mathbf{R}$, we obtain the following series expansion:

$$\mathbf{R}=(1-c)\sum_{t=0}^{\infty}\tfrac{c^t}{2^t}\sum_{a=0}^{t}\tbinom{t}{a}\mathbf{P}^a(\mathbf{P}^\top)^{t-a}.$$

The q-th column vector $\mathbf{r}=\mathbf{R}\mathbf{q}$ can be represented as

$$\mathbf{r}=(1-c)\sum_{t=0}^{\infty}\tfrac{c^t}{2^t}\sum_{a=0}^{t}\tbinom{t}{a}\mathbf{P}^a(\mathbf{P}^\top)^{t-a}\mathbf{q}.$$

We truncate the series to approximate the original proximity. Let η denote the truncated length and $\hat{\mathbf{r}}$ denote the truncated proximity. We have that

$$\hat{\mathbf{r}}=(1-c)\sum_{t=0}^{\eta}\tfrac{c^t}{2^t}\sum_{a=0}^{t}\tbinom{t}{a}\mathbf{P}^a(\mathbf{P}^\top)^{t-a}\mathbf{q}.$$

Only the meeting paths of length $\{a,t\}$ where $t\leq\eta$ are counted. We replace the index a with $(t-a)$ and have that

$$\hat{\mathbf{r}}=(1-c)\sum_{t=0}^{\eta}\tfrac{c^t}{2^t}\sum_{a=0}^{t}\tbinom{t}{a}\mathbf{P}^{t-a}(\mathbf{P}^\top)^{a}\mathbf{q}.$$

We interchange the index t and a and have that

$$\hat{\mathbf{r}}=(1-c)\sum_{a=0}^{\eta}\sum_{t=a}^{\eta}\tfrac{c^t}{2^t}\tbinom{t}{a}\mathbf{P}^{t-a}(\mathbf{P}^\top)^{a}\mathbf{q}.$$

D.4 The second-order SimRank*

Recall that the second-order SimRank* is defined in Equation (4) and shown as follows.

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

We truncate the series to approximate the original proximity. Let η denote the truncated length and $\hat{r}_{i,j}$ denote the truncated proximity. We have that

$$\hat{r}_{i,j}=(1-c)\sum_{t=0}^{\eta}\tfrac{c^t}{2^t}\sum_{a=0}^{t}\tbinom{t}{a}\mathbb{M}[\Phi_{i,j}^{a,t}].$$

Only the meeting paths of length $\{a,t\}$ where $t\leq\eta$ are counted. We replace the index a with $(t-a)$ and have that

$$\hat{r}_{i,j}=(1-c)\sum_{t=0}^{\eta}\tfrac{c^t}{2^t}\sum_{a=0}^{t}\tbinom{t}{a}\mathbb{M}[\Phi_{i,j}^{t-a,t}].$$

We interchange the index t and a and have that

$$\hat{r}_{i,j}=(1-c)\sum_{a=0}^{\eta}\sum_{t=a}^{\eta}\tfrac{c^t}{2^t}\tbinom{t}{a}\mathbb{M}[\Phi_{i,j}^{t-a,t}] \tag{14}$$

Based on Lemma 2, we have the following equations.

$$\mathbb{M}[\Phi_{i,j}^{t-a,t}]=\begin{cases}\mathbf{I}_{i,j}, & \text{if } 0=a=t,\\ [\mathbf{H}\mathbf{M}^{t-1}\mathbf{E}]_{i,j}, & \text{if } 0=a<t,\\ [\mathbf{E}^\top(\mathbf{M}^\top)^{t-1}\mathbf{H}^\top]_{i,j}, & \text{if } 0<a=t,\\ [\mathbf{H}\mathbf{M}^{t-a-1}\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{a-1}\mathbf{H}^\top]_{i,j}, & \text{if } 0<a<t.\end{cases}$$

Let $\hat{\mathbf{R}}$ denote the approximate proximity matrix with $[\hat{\mathbf{R}}]_{i,j}=\hat{r}_{i,j}$. Plugging the above equations into Equation (14), we have the following equation.

$$\begin{aligned}\hat{\mathbf{R}}=&(1-c)\mathbf{I}+(1-c)\textstyle\sum_{t=1}^{\eta}\frac{c^t}{2^t}\mathbf{H}\mathbf{M}^{t-1}\mathbf{E}\\&+(1-c)\textstyle\sum_{t=1}^{\eta}\frac{c^t}{2^t}\mathbf{E}^\top(\mathbf{M}^\top)^{t-1}\mathbf{H}^\top\\&+(1-c)\textstyle\sum_{a=1}^{\eta-1}\sum_{t=a+1}^{\eta}\frac{c^t}{2^t}\binom{t}{a}\mathbf{H}\mathbf{M}^{t-a-1}\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{a-1}\mathbf{H}^\top\end{aligned}$$

The q-th column vector $\hat{\mathbf{r}}=\hat{\mathbf{R}}\mathbf{q}$ can be represented as

$$\begin{aligned}\hat{\mathbf{r}}=&(1-c)\mathbf{q}+(1-c)\textstyle\sum_{t=1}^{\eta}\frac{c^t}{2^t}\mathbf{H}\mathbf{M}^{t-1}\mathbf{E}\mathbf{q}\\&+(1-c)\textstyle\sum_{t=1}^{\eta}\frac{c^t}{2^t}\mathbf{E}^\top(\mathbf{M}^\top)^{t-1}\mathbf{H}^\top\mathbf{q}\\&+(1-c)\textstyle\sum_{a=1}^{\eta-1}\sum_{t=a+1}^{\eta}\frac{c^t}{2^t}\binom{t}{a}\mathbf{H}\mathbf{M}^{t-a-1}\mathbf{E}\mathbf{E}^\top(\mathbf{M}^\top)^{a-1}\mathbf{H}^\top\mathbf{q}\end{aligned}$$

E The Monte Carlo method

E.1 The second-order random walk with restart

Lemma 8 is needed in the proof of Theorem 6.

Lemma 8 $\mathbb{E}[\mathbb{S}_i]=r_i$

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

Table 9 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

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

$$\mathbb{E}[\mathbb{S}_i] = \sum_{a=0}^{\infty} \mathbb{P}[\mathbb{A}=a] \cdot \mathbb{E}[\mathbb{S}_i | \mathbb{A}=a], \tag{15}$$

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 9 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 $\underline{\mathbb{M}}[\rho] = p_{z_0,z_1} \cdot \prod_{t=1}^{a-1} p_{z_{t-1},z_t,z_{t+1}}$. 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

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

Plugging this into Equation (15), we have that

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

This completes the proof of Lemma 8. □

Table 10 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)	$z_{2a} = i$	δ
		$z_{2a} \neq i$	0
	fail to sample a path of length a starting from node q (bSuccess = false)	–	0
$\eta < a$	–	–	0

E.2 The first-order SimRank

Lemma 9 is needed in the proofs of Theorems 9 and 10.

Lemma 9 $\mathbb{E}[\mathbb{R}_i] = \hat{r}_i$ *and* $\mathbb{E}[\mathbb{R}_i^2] \le n r_i$

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 $a (0 \le a \le \eta)$ 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 $\{0, \delta\}$.

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

$$\mathbb{E}[\mathbb{R}_i] = \sum_{a=0}^{\infty} \mathbb{P}[\mathbb{A}=a] \cdot \mathbb{E}[\mathbb{R}_i | \mathbb{A}=a], \tag{16}$$

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 10 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 $\phi: q=z_0\rightarrow\cdots\rightarrow z_a\leftarrow\cdots\leftarrow z_{2a}$ 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

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

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 $0\leq a\leq\eta$, we have

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

If $\eta<a$, we have $\mathbb{E}[\mathbb{R}_i|\mathbb{A}=a]=0$. Plugging this into Equation (16), we have that

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

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

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

$$\mathbb{E}[\mathbb{R}_i^2]=\textstyle\sum_{a=0}^{\infty}\mathbb{P}[\mathbb{A}=a]\cdot\mathbb{E}[\mathbb{R}_i^2|\mathbb{A}=a] \tag{17}$$

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 $\delta=[\mathbf{X}]_{z_a,a}/[\mathbf{X}]_{z_{2a},0}\leq n$. Thus, if $0\leq a\leq\eta$, we have

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

If $\eta<a$, we have $\mathbb{E}[\mathbb{R}_i^2|\mathbb{A}=a]=0$. Plugging this into Equation (17), we have that

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

This completes the proof of Lemma 9. □

Theorem 22 is needed in the proofs of Theorems 10, 12, 14, and 16. Theorem 22 is based on Theorems 2.8 and 2.9 in [5].

Theorem 22 [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., $-\alpha\leq\mathbb{U}_d\leq\beta$ 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*

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

Table 11 The sample space when the length a is given

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

E.3 The second-order SimRank

Lemma 10 is needed in the proofs of Theorems 11 and 12.

Lemma 10 $\mathbb{E}[\mathbb{S}_i]=\hat{r}_i$ *and* $\mathbb{E}[\mathbb{S}_i^2]\leq\kappa mr_i$

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 $a\,(0\leq a\leq\eta)$ and successfully samples a meeting path of length $\{a,2a\}$ starting from the query node q.
2) The algorithm generates a random number $a\,(0\leq a\leq\eta)$ 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 $\mathbb{S}_i$, which is the set of real values $\{0,\delta\}$.

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

$$\mathbb{E}[\mathbb{S}_i]=\textstyle\sum_{a=0}^{\infty}\mathbb{P}[\mathbb{A}=a]\cdot\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a]\,, \tag{18}$$

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 11 shows the sample space given the length a. Given the length $a\,(0\leq a\leq\eta)$, if the algorithm successfully samples a meeting path of length $\{a,2a\}$ between nodes q and i, the random variable $\mathbb{S}_i=\delta$; otherwise, $\mathbb{S}_i=0$. The incremental value δ changes for different sampled meeting paths. Let $\underline{\mathbb{M}}[\phi]$ represent the probability of successfully sampling a meeting path ϕ given the length a. Since $\mathbb{S}_i=\delta$ if and only if the algorithm successfully samples a meeting path ϕ of length $\{a,2a\}$ between nodes q and i, the conditional expectation $\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a]$ can be written as

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

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 $\underline{\mathbb{M}}[\rho_1]=p_{z_0,z_1}\prod_{t=2}^{a}p_{z_{t-2},z_{t-1},z_t}$. The probability of sampling the second half is

$$\begin{aligned}\underline{\mathbb{M}}[\rho_2] &= \frac{1}{|I_{z_a}|}\prod_{t=a+2}^{2a}\left(p_{z_t,z_{t-1},z_{t-2}}\cdot\frac{[\mathbf{Y}]_{(z_t,z_{t-1}),2a-t}}{[\mathbf{Y}]_{(z_{t-1},z_{t-2}),2a-t+1}}\right)\\ &= \frac{1}{|I_{z_a}|}\cdot\frac{[\mathbf{Y}]_{(z_{2a},z_{2a-1}),0}}{[\mathbf{Y}]_{(z_{a+1},z_a),a-1}}\prod_{t=a+2}^{2a}p_{z_t,z_{t-1},z_{t-2}}\end{aligned}$$

Note that the incremental value δ is as follows

$$\delta=|I_{z_a}|\cdot p_{z_{2a},z_{2a-1}}\cdot\frac{[\mathbf{Y}]_{(z_{a+1},z_a),a-1}}{[\mathbf{Y}]_{(z_{2a},z_{2a-1}),0}}.$$

Thus, we have the following equation

$$\delta\cdot\underline{\mathbb{M}}[\rho_2]=p_{z_{2a},z_{2a-1}}\prod_{t=a+2}^{2a}p_{z_t,z_{t-1},z_{t-2}}.$$

We can see that the right hand side is equal to the probability of visiting path ρ_2, i.e., $\delta\cdot\underline{\mathbb{M}}[\rho_2]=\mathbb{M}[\rho_2]$. The probability of sampling the meeting path ϕ is $\underline{\mathbb{M}}[\phi]=\underline{\mathbb{M}}[\rho_1]\cdot\underline{\mathbb{M}}[\rho_2]$. Thus the probabilities of sampling and visiting a meeting path ϕ have a relationship, i.e., $\delta\cdot\underline{\mathbb{M}}[\phi]=\mathbb{M}[\phi]$. Thus, if $0\le a\le\eta$, we have

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

If $\eta<a$, we have $\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a]=0$. Plugging this into Equation (18), we have that

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

where $\hat{r}_i$ is the truncated second-order SimRank proximity value.

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

$$\mathbb{E}[\mathbb{S}_i^2]=\textstyle\sum_{a=0}^{\infty}\mathbb{P}[\mathbb{A}=a]\cdot\mathbb{E}[\mathbb{S}_i^2|\mathbb{A}=a] \tag{19}$$

Since $[\mathbf{Y}]_{(z_{a+1},z_a),a-1}\in[0,1]$, $[\mathbf{Y}]_{(z_{2a},z_{2a-1}),0}=\frac{1}{m}$, and $|I_{z_a}|\le\kappa$, where m is the number of edges in the graph and κ is the maximum in-degree, we have that $\delta\le\kappa m$. Thus, if $0\le a\le\eta$, we have

$$\begin{aligned}\mathbb{E}[\mathbb{S}_i^2|\mathbb{A}=a] &= \textstyle\sum_{\phi\in\Phi_{q,i}^{a,2a}}\delta^2\underline{\mathbb{M}}[\phi]\\ &= \textstyle\sum_{\phi\in\Phi_{q,i}^{a,2a}}\delta\,\mathbb{M}[\phi]\le\kappa m\mathbb{M}[\Phi_{q,i}^{a,2a}]\end{aligned}$$

If $\eta<a$, we have $\mathbb{E}[\mathbb{S}_i^2|\mathbb{A}=a]=0$. Plugging this into Equation (19), we have that

$$\mathbb{E}[\mathbb{S}_i^2]\le(1-c)\textstyle\sum_{a=0}^{\infty}c^a\kappa m\mathbb{M}[\Phi_{q,i}^{a,2a}]=\kappa m r_i$$

This completes the proof of Lemma 10. □

Table 12 The sample space when the length $\{a,b\}$ is given

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

E.4 The first-order SimRank*

The proof of Theorem 13 is as follows.

Proof In Algorithm 14, if we successfully sample a path ending at node i, we will increase $\tilde{r}_i$ by δ; otherwise, $\tilde{r}_i$ is unchanged. For different sampled paths, the corresponding value δ may be different. Let $\mathbb{R}_i^{(d)}$ be a random variable denoting the incremental value of $\tilde{r}_i$ at the d-th iteration (lines $3\sim7$). Random variables $\mathbb{R}_i^{(1)},\mathbb{R}_i^{(2)},\cdots,\mathbb{R}_i^{(\pi)}$ are independent and identically distributed. Let $\mathbb{R}_i$ be a random variable following the same distribution as $\mathbb{R}_i^{(d)}$'s. Lemma 11 shows that the expected value of $\mathbb{R}_i$ equals the truncated proximity $\hat{r}_i$, i.e., $\mathbb{E}[\mathbb{R}_i]=\hat{r}_i$.

Lemma 11 $\mathbb{E}[\mathbb{R}_i]=\hat{r}_i$ *and* $\mathbb{E}[\mathbb{R}_i^2]\le n r_i$

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

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

$$\mathbb{E}[\mathbb{R}_i]=\sum_{b=0}^{\infty}\sum_{a=0}^{b}\mathbb{P}[\mathbb{B}=b]\cdot\mathbb{P}[\mathbb{A}=a]\cdot\mathbb{E}[\mathbb{R}_i|\mathbb{A}=a,\mathbb{B}=b], \tag{20}$$

where the random variable $\mathbb{B}$ representing the length b follows the geometric distribution $\mathbb{P}[\mathbb{B}=b]=(1-c)\cdot c^b$ and the random variable $\mathbb{A}$ representing the length a follows the binomial distribution $\mathbb{P}[\mathbb{A}=a]=\binom{b}{a}/2^b$. Next, we consider the conditional expectation $\mathbb{E}[\mathbb{R}_i|\mathbb{A}=a,\mathbb{B}=b]$ of $\mathbb{R}_i$ given the event $\mathbb{A}=a$ and $\mathbb{B}=b$.

Table 12 shows the sample space given the length $\{a,b\}$. Given the length $b(0\le b\le\eta)$ and $a(0\le a\le b)$, if the algorithm successfully samples a meeting path of length $\{a,b\}$ between nodes q and i, the random variable $\mathbb{R}_i=\delta$; otherwise, $\mathbb{R}_i=0$. Note that $\delta=[\mathbf{X}]_{z_a,b-a}/[\mathbf{X}]_{z_b,0}$ changes for different sampled meeting paths. Let $\underline{\mathbb{P}}[\phi]$ represent the probability of successfully sampling a meeting path ϕ given the length $\{a,b\}$. Since $\mathbb{R}_i=\delta$ if and only if the algorithm successfully samples a meeting path ϕ of length $\{a,b\}$ between nodes q and i, the conditional expectation $\mathbb{E}[\mathbb{R}_i|\mathbb{A}=a,\mathbb{B}=b]$ can be written as

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

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

Consider the probability of successfully sampling a meeting path $\phi: q=z_0\to\cdots\to z_a\leftarrow\cdots\leftarrow z_b$ given the length $\{a,b\}$. 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

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

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 $\delta=[\mathbf{X}]_{z_a,b-a}/[\mathbf{X}]_{z_b,0}$. 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 $0\le b\le\eta$, we have

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

If $\eta<b$, we have $\mathbb{E}[\mathbb{R}_i|\mathbb{A}=a,\mathbb{B}=b]=0$. Plugging this into Equation (20), we have that

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

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

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

$$\mathbb{E}[\mathbb{R}_i^2]=\sum_{b=0}^{\infty}\sum_{a=0}^{b}\mathbb{P}[\mathbb{B}=b]\cdot\mathbb{P}[\mathbb{A}=a]\cdot\mathbb{E}[\mathbb{R}_i^2|\mathbb{A}=a,\mathbb{B}=b]\quad(21)$$

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

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

If $\eta<b$, we have $\mathbb{E}[\mathbb{R}_i^2|\mathbb{A}=a,\mathbb{B}=b]=0$. Plugging this into Equation (21), we have that

$$\mathbb{E}[\mathbb{R}_i^2]\le(1-c)\textstyle\sum_{b=0}^{\infty}\frac{c^b}{2^b}\sum_{a=0}^{b}\binom{b}{a}n\mathbb{P}[\Phi_{q,i}^{a,b}]=nr_i$$

This completes the proof of Lemma 11. □

Let $\overline{\mathbb{R}}_i=\frac{1}{\pi}\sum_{d=1}^{\pi}\mathbb{R}_i^{(d)}$ be the sample average, which represents the estimated proximity $\tilde{r}_i$. By the law of large numbers, if the sample size $\pi\to\infty$, $\overline{\mathbb{R}}_i$ converges to the expected value $\mathbb{E}[\mathbb{R}_i]=\hat{r}_i$. This completes the proof of Theorem 13. □

The proof of Theorem 14 is as follows.

Proof Following the notations defined in the proof of Theorem 13, random variables $\mathbb{R}_i^{(1)},\mathbb{R}_i^{(2)},\cdots,\mathbb{R}_i^{(\pi)}$ are independent and bounded by interval $[0,\delta]\subseteq[0,n]$. Lemma 11 shows that the expected value of $\mathbb{R}_i^2$ is bounded from above by nr_i, i.e., $\mathbb{E}[\mathbb{R}_i^2]\le nr_i$. Thus, we have that $\sum_{d=1}^{\pi}\mathbb{E}[(\mathbb{R}_i^{(d)})^2]\le\pi nr_i$. By Theorem 22 in Appendix E.2, we can prove this theorem. □

E.5 The second-order SimRank*

The proof of Theorem 15 is as follows.

Proof In Algorithm 16, if we successfully sample a path ending at node i, we will increase $\tilde{r}_i$ by δ; otherwise, $\tilde{r}_i$ is unchanged. For different sampled paths, the corresponding value δ may be different. Let $\mathbb{S}_i^{(d)}$ be a random variable denoting the incremental value of $\tilde{r}_i$ at the d-th iteration (lines 3∼7). Random variables $\mathbb{S}_i^{(1)},\mathbb{S}_i^{(2)},\cdots,\mathbb{S}_i^{(\pi)}$ are independent and identically distributed. Let $\mathbb{S}_i$ be a random variable following the same distribution as $\mathbb{S}_i^{(d)}$'s. Lemma 12 shows that the expected value of $\mathbb{S}_i$ equals the truncated proximity $\hat{r}_i$, i.e., $\mathbb{E}[\mathbb{S}_i]=\hat{r}_i$.

Lemma 12 $\mathbb{E}[\mathbb{S}_i]=\hat{r}_i$ *and* $\mathbb{E}[\mathbb{S}_i^2]\le\kappa mr_i$

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

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

$$\mathbb{E}[\mathbb{S}_i]=\sum_{b=0}^{\infty}\sum_{a=0}^{b}\mathbb{P}[\mathbb{B}=b]\cdot\mathbb{P}[\mathbb{A}=a]\cdot\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a,\mathbb{B}=b],\quad(22)$$

Table 13 The sample space when the length $\{a,b\}$ is given

$\{a,b\}$	sample space		$\mathbb{S}_i$
	bSuccess	node z_b	
$0\le b\le\eta$	successfully sample a path of length $\{a,b\}$ starting from node q (bSuccess=true)	$z_b=i$	δ
		$z_b\neq i$	0
$0\le a\le b$	fail to sample a path of length $\{a,b\}$ starting from node q (bSuccess = false)	–	0
$\eta<b$	–	–	0

where the random variable $\mathbb{B}$ representing the length b follows the geometric distribution $\mathbb{P}[\mathbb{B}=b]=(1-c)\cdot c^b$ and the random variable $\mathbb{A}$ representing the length a follows the binomial distribution $\mathbb{P}[\mathbb{A}=a]=\binom{b}{a}/2^b$. Next, we consider the conditional expectation $\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a,\mathbb{B}=b]$ of $\mathbb{S}_i$ given the event $\mathbb{A}=a$ and $\mathbb{B}=b$.

Table 13 shows the sample space given the length $\{a,b\}$. Given the length $b(0\le b\le\eta)$ and $a(0\le a\le b)$, if the algorithm successfully samples a meeting path of length $\{a,b\}$ between nodes q and i, the random variable $\mathbb{S}_i=\delta$; otherwise, $\mathbb{S}_i=0$. The incremental value δ changes for different sampled meeting paths. Let $\underline{\mathbb{M}}[\phi]$ represent the probability of successfully sampling a meeting path ϕ given the length $\{a,b\}$. Since $\mathbb{S}_i=\delta$ if and only if the algorithm successfully samples a meeting path ϕ of length $\{a,b\}$ between nodes q and i, the conditional expectation $\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a,\mathbb{B}=b]$ can be written as

$$\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a,\mathbb{B}=b]=\sum\nolimits_{\phi\in\Phi^{a,b}_{q,i}}\delta\cdot\underline{\mathbb{M}}[\phi],$$

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

Consider the probability of successfully sampling a meeting path $\phi: q=z_0\to\cdots\to z_a\leftarrow\cdots\leftarrow z_b$ given the length $\{a,b\}$. The probability of sampling the first half is $\underline{\mathbb{M}}[\rho_1]=p_{z_0,z_1}\prod_{t=2}^{a}p_{z_{t-2},z_{t-1},z_t}$. The probability of sampling the second half is

$$\begin{aligned}\underline{\mathbb{M}}[\rho_2]&=\frac{1}{|I_{z_a}|}\prod_{t=a+2}^{b}\Big(p_{z_t,z_{t-1},z_{t-2}}\cdot\frac{[\mathbf{Y}]_{(z_t,z_{t-1}),b-t}}{[\mathbf{Y}]_{(z_{t-1},z_{t-2}),b-t+1}}\Big)\\&=\frac{1}{|I_{z_a}|}\cdot\frac{[\mathbf{Y}]_{(z_b,z_{b-1}),0}}{[\mathbf{Y}]_{(z_{a+1},z_a),b-a-1}}\prod_{t=a+2}^{b}p_{z_t,z_{t-1},z_{t-2}}\end{aligned}$$

Note that the incremental value δ is as follows

$$\delta=|I_{z_a}|\cdot p_{z_b,z_{b-1}}\cdot\frac{[\mathbf{Y}]_{(z_{a+1},z_a),b-a-1}}{[\mathbf{Y}]_{(z_b,z_{b-1}),0}}.$$

Thus, we have the following equation

$$\delta\cdot\underline{\mathbb{M}}[\rho_2]=p_{z_b,z_{b-1}}\prod_{t=a+2}^{b}p_{z_t,z_{t-1},z_{t-2}}.$$

We can see that the right hand side is equal to the probability of visiting path ρ_2, i.e., $\delta\cdot\underline{\mathbb{M}}[\rho_2]=\mathbb{M}[\rho_2]$. The probability of sampling the meeting path ϕ is $\underline{\mathbb{M}}[\phi]=\underline{\mathbb{M}}[\rho_1]\cdot\underline{\mathbb{M}}[\rho_2]$. Thus the probabilities of sampling and visiting a meeting path ϕ have a relationship, i.e., $\delta\cdot\underline{\mathbb{M}}[\phi]=\mathbb{M}[\phi]$. Thus, if $0\le b\le\eta$, we have

$$\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a,\mathbb{B}=b]=\sum\nolimits_{\phi\in\Phi^{a,b}_{q,i}}\mathbb{M}[\phi]=\mathbb{M}[\Phi^{a,b}_{q,i}]$$

If $\eta<b$, we have $\mathbb{E}[\mathbb{S}_i|\mathbb{A}=a,\mathbb{B}=b]=0$. Plugging this into Equation (22), we have that

$$\mathbb{E}[\mathbb{S}_i]=(1-c)\sum\nolimits_{b=0}^{\infty}\tfrac{c^b}{2^b}\sum\nolimits_{a=0}^{b}\tbinom{b}{a}\mathbb{M}[\Phi^{a,b}_{q,i}]=\hat{r}_i\,,$$

where $\hat{r}_i$ is the truncated second-order SimRank* proximity value.

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

$$\mathbb{E}[\mathbb{S}_i^2]=\sum_{b=0}^{\infty}\sum_{a=0}^{b}\mathbb{P}[\mathbb{B}=b]\cdot\mathbb{P}[\mathbb{A}=a]\cdot\mathbb{E}[\mathbb{S}_i^2|\mathbb{A}=a,\mathbb{B}=b]\quad(23)$$

Since $[\mathbf{Y}]_{(z_{a+1},z_a),b-a-1}\in[0,1]$, $[\mathbf{Y}]_{(z_b,z_{b-1}),0}=\frac{1}{m}$, and $|I_{z_a}|\le\kappa$, where m is the number of edges in the graph and κ is the maximum in-degree, we have that $\delta\le\kappa m$. Thus, if $0\le a\le\eta$, we have

$$\begin{aligned}\mathbb{E}[\mathbb{S}_i^2|\mathbb{A}=a,\mathbb{B}=b]&=\sum\nolimits_{\phi\in\Phi^{a,b}_{q,i}}\delta^2\,\underline{\mathbb{M}}[\phi]\\&=\sum\nolimits_{\phi\in\Phi^{a,b}_{q,i}}\delta\,\mathbb{M}[\phi]\le\kappa m\mathbb{M}[\Phi^{a,b}_{q,i}]\end{aligned}$$

If $\eta<a$, we have $\mathbb{E}[\mathbb{S}_i^2|\mathbb{A}=a,\mathbb{B}=b]=0$. Plugging this into Equation (23), we have that

$$\mathbb{E}[\mathbb{S}_i^2]\le(1-c)\sum\nolimits_{b=0}^{\infty}\tfrac{c^b}{2^b}\sum\nolimits_{a=0}^{b}\tbinom{b}{a}\kappa m\mathbb{M}[\Phi^{a,b}_{q,i}]=\kappa m r_i\,,$$

This completes the proof of Lemma 12. □

Let $\overline{\mathbb{S}}_i=\frac{1}{\pi}\sum_{d=1}^{\pi}\mathbb{S}_i^{(d)}$ be the sample average, which represents the estimated proximity $\tilde{r}_i$. By the law of large numbers, if the sample size $\pi\to\infty$, $\overline{\mathbb{S}}_i$ converges to the expected value $\mathbb{E}[\mathbb{S}_i]=\hat{r}_i$. This completes the proof of Theorem 15. □

The proof of Theorem 16 is as follows.

Proof Following the notations defined in the proof of Theorem 15, random variables $\mathbb{S}_i^{(1)},\mathbb{S}_i^{(2)},\cdots,\mathbb{S}_i^{(\pi)}$ are independent and bounded by interval $[0,\delta]\subseteq[0,\kappa m]$. Lemma 12 shows that the expected value of $\mathbb{S}_i^2$ is bounded from above by $\kappa m r_i$, i.e., $\mathbb{E}[\mathbb{S}_i^2]\le\kappa m r_i$. Thus, we have that $\sum_{d=1}^{\pi}\mathbb{E}[(\mathbb{S}_i^{(d)})^2]\le\pi\kappa m r_i$. By Theorem 22 in Appendix E.2, we can prove this theorem. □