To Reviewer #1. We appreciate your positive feedback and will revise our presentation accordingly. Our bounds can help hyper-parameter selection in graph representation learning (a running example about BlogCatalog can be found in our response to Reviewer #4). Prior to this work, the walk length of DeepWalk has to be selected by cross-validation.

To Reviewer #2. Thank you for your comments. We appreciate your views and we would like to clarify a few points.

[The Paper Framing] Our original intention is to analyze a graph representation learning algorithm, DeepWalk, which involves sampling random walks from a graph and counting the vertex co-occurrence matrix. Indeed, the generalized matrix Chernoff bounds are powerful and its proof is the most challenging part. So it turns out that we solved a more fundamental theory problem when studying the particular application. We are open to reframing the work as "Matrix Chernoff Bounds for Ergodic Markov Chains and its Application to Co-occurrence Matrices".

[Hidden Markov Models (HMM)] For a HMM, let us denote Y and X to be the space of observable states and hidden states, respectively. A HMM can be model by a Markov chain $\boldsymbol{P}'$ on $Y \times X$ such that $P'(y_{t+1}, x_{t+1}|y_t, x_t) = P(y_{t+1}|x_{t+1})P(x_{t+1}|x_t)$, where $P(y_{t+1}|x_{t+1})$ is the emission probability and $P(x_{t+1}|x_t)$ is the hidden state transition probability. If the co-occurrence matrix is defined only on the observable state space Y, then applying a similar proof like our Theorem 1 shows that one needs a trajectory of length $O(\tau(\log|Y| + \log\tau)/\epsilon^2)$ to achieve error bound ϵ, where τ is the mixing time of Markov chain $\boldsymbol{P}'$ and the space of hidden states X could be large. Moreover, the mixing time of Markov chain $\boldsymbol{P}'$ is bounded by the mixing time of the Markov chain on the hidden state space (i.e., $P(x_{t+1}|x_t)$).

To Reviewer #3. Thank you for your comments. We appreciate your views and we would like to clarify a few points.

[The Tightness of the Bound] The bound on co-occurrence matrices may not be tight. In our proof, we need to partition the chain $\boldsymbol{Q}$ into $\tau(\boldsymbol{Q})$ groups and then combine them with union bound, which probably gives a loose bound. As we have mentioned in 'Conclusion and Future Work' section (Sec. 6), it would be interesting to shave off the leading factor τ in the bound, as the mixing time τ could be large for some Markov chains.

[Regarding Initial Distribution] Thanks a lot for pointing out this! In our latest version, we have allowed the Markov chain to start from an arbitrary initial distribution $\boldsymbol{\phi}$ rather than the stationary distribution $\boldsymbol{\pi}$. And there will be an additional term measuring the distance between $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$ in our new bound.

[Markov Chains with Continuous States] For infinite or continuous Markov chain, it appears to us this is a non-trivial extension and is thus beyond the scope of our paper (but certainly very interesting direction for future study). Technically speaking proving such results requires a non-trivial extension of the matrix bound (Theorem 3 in our paper), and this requires a lot more work and not the main purpose of current paper.

To Reviewer #4. Thank you for your comments. We appreciate your views and we would like to clarify a few points.

[Regarding Mixing Time] We agree that the mixing time of Markov chains are usually unknown in advance. However, it can be estimated statistically, e.g., [41]. Empirically, many real-world networks have the rapid mixing property.

[The BlogCatalog Experiment from Qiu et al.] Our bounds on trajectory length L in Theorem 1 (with explicit constant) is $L \geq 576(\tau + T)(3\log n + \log(\tau + T))/\epsilon^2 + T$. The error bound ϵ might be chosen in the range of $[0.1, 0.01]$, which corresponds to L in the range of $[8.4 \times 10^7, 8.4 \times 10^9]$. To verify that is a meaningful range for tuning L, we enumerate trajectory length L from $\{10^4, \cdots, 10^{10}\}$, estimate the co-occurrence matrix with the single trajectory sampled from BlogCatalog, convert the co-occurrence matrix to the one required by NetMF, and factorize it with SVD. For node classification task, the micro-F1 when training ratio is 50% is

Length L	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}	NetMF
Micro-F1 (%)	15.21	18.31	26.99	33.85	39.12	41.28	41.58	**41.82**

. As we can see, it is reasonable to choose L in the predicted range. Due to page limit of author responses, we have to put more detailed results in our next version.

To Reviewer #5. Thank you for pointing us to relevant literature and techniques that we were not aware of before: our starting point was the random walk based graph embedding methods, and it's great to know that there are many more techniques that can be used to analyze them.

[Prior Work] Claim 1 in our paper (of how the T-step walk is itself a Markov chain) is indeed a generalization of Lemma 6.1 from [42], which discusses the special case of $T = 1$. We will cite and discuss all the related papers you mentioned ([42-45]), as well as how the bounds formally relate, in our next version.

[Blocking Techniques] Thank you for pointing us to this paper (Hsu et al. [43]) and the blocking techniques based on dividing the walk into nearly independent blocks of lenth around mixing time. We agree that this technique can also be used to analyze the convergence of co-occurrence matrices and much more: Garg et al. [11] was just one way to set up the analysis, and we will more clearly indicate this after adding the appropriate comparisons and references.

[Regarding Initial Distribution] This question is also raised by Reviewer #3. As we mentioned above, we have allowed the Markov chain to start from an arbitrary initial distribution in our latest version.