NEURAL INFORMATION PROCESSING SYSTEMS

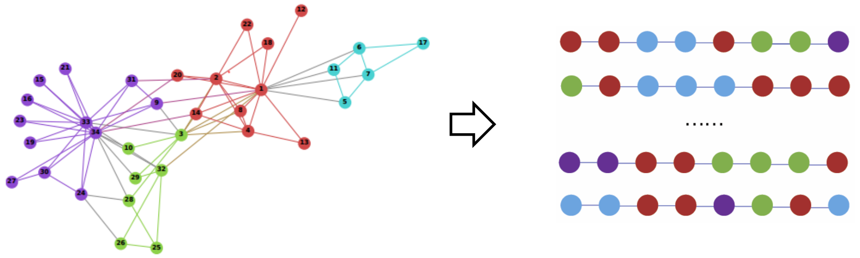
A Matrix Chernoff Bound for Markov Chains and its Application to Co-occurrence Matrices

Jiezhong Qiu, Chi Wang, Ben Liao, Richard Peng, Jie Tang

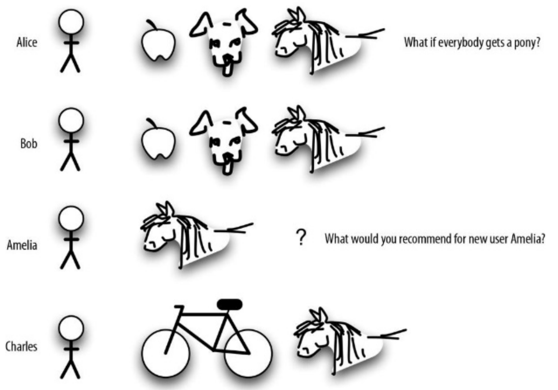
The Application to Co-occurrence Matrices

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

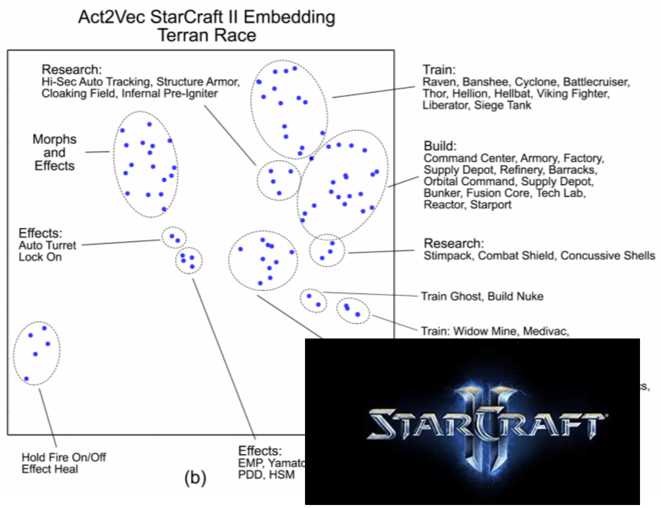
NLP
(LDA, Word2vec, Glove)



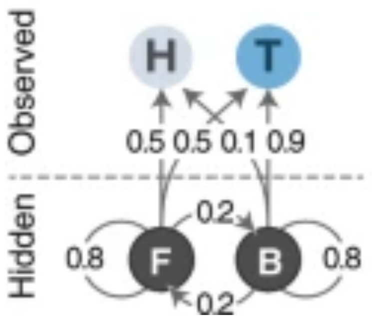
Graph Learning
(DeepWalk, node2vec, metapath2vec)



Recommendation System
(Pin2Vec, Item2vec)



Reinforcement Learning
(Act2Vec)

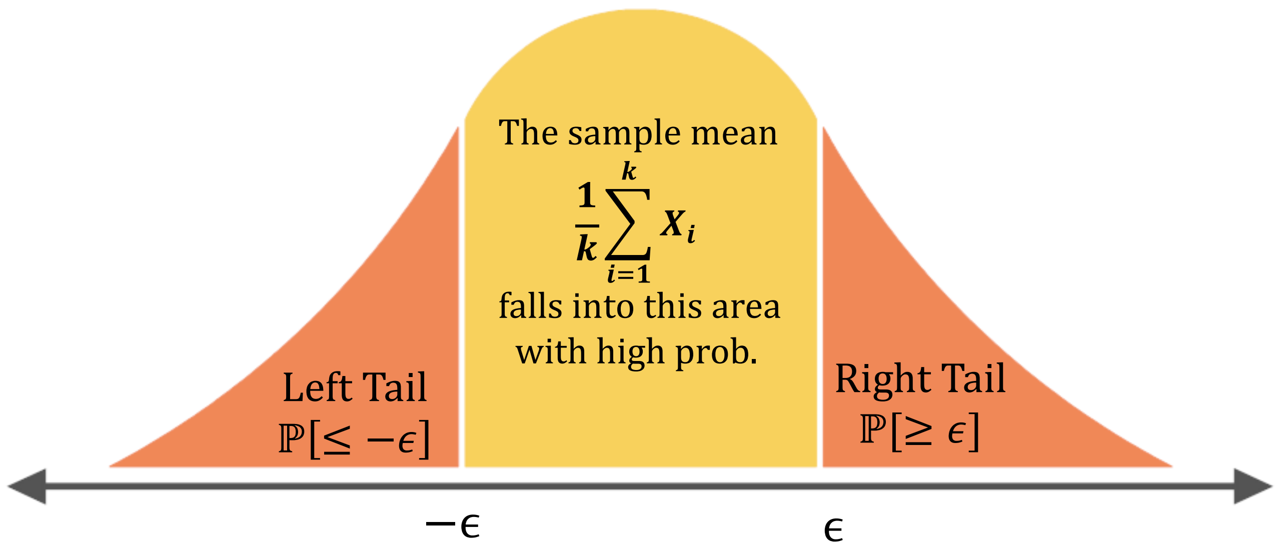


Hidden Markov Models
(Emission Co-occurrence)

Chernoff Bounds

Theorem (Chernoff Bound, 1952): If $X_1, X_2, \cdots, X_k$ are independent zero-mean scaler-valued random variables with $|X_i| \leq 1$. Then for $\epsilon \in (0, 1)$

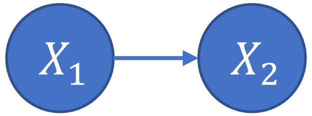
$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{i=1}^{k} X_i\right| \geq \epsilon\right) \leq 2\exp(-k\epsilon^2/4)$$



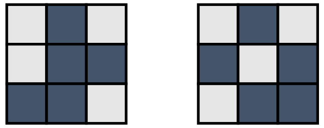
A Matrix Chernoff Bound for Markov Chains

~~Independence~~
Markov Dependence

$P(X_2|X_1)$



~~Scalar-valued Random Variables~~
Matrix-valued Random Variables



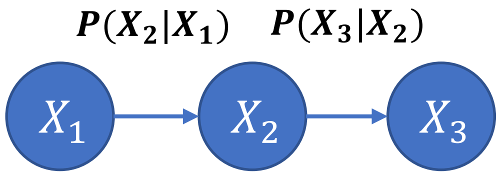
$f(X_1)$ $f(X_2)$

Sample Mean Matrix

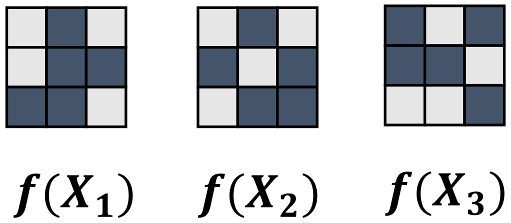
$$\frac{1}{2}(f(X_1) + f(X_2))$$

A Matrix Chernoff Bound for Markov Chains

~~Independence~~
Markov Dependence



~~Scalar-valued Random Variables~~
Matrix-valued Random Variables

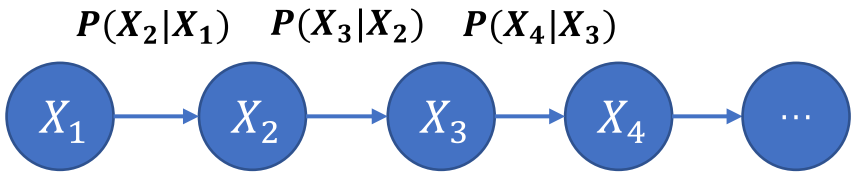


Sample Mean Matrix

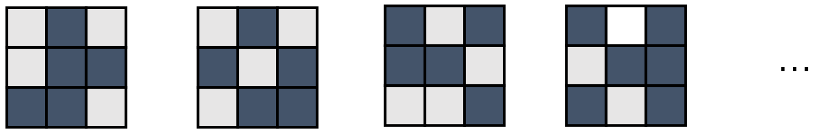
$$\frac{1}{3}(f(X_1) + f(X_2) + f(X_3))$$

A Matrix Chernoff Bound for Markov Chains

~~Independence~~
Markov Dependence



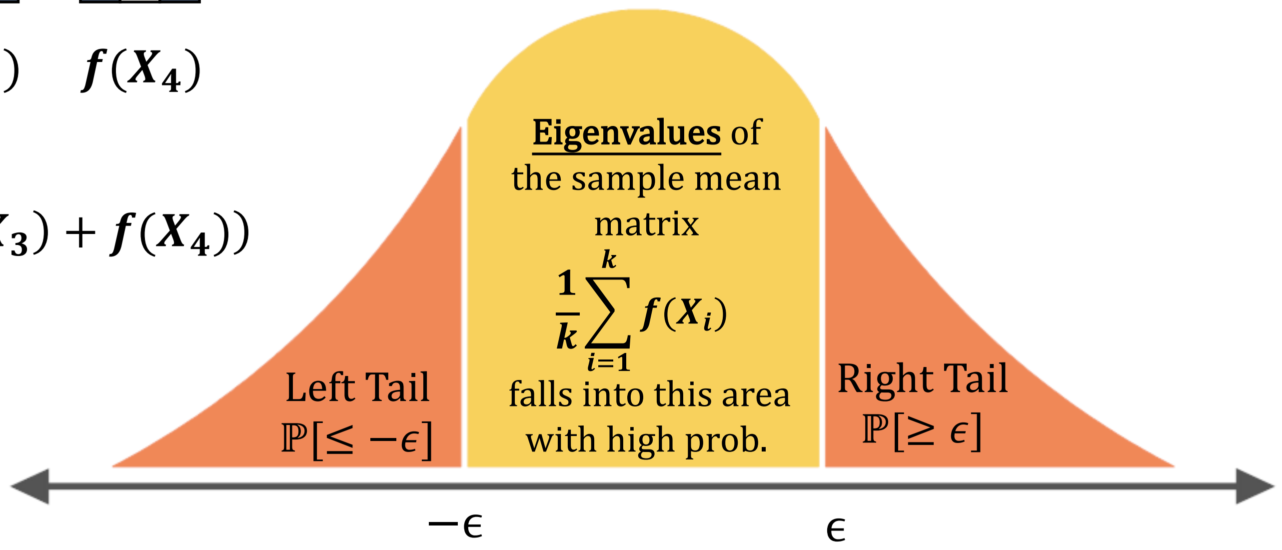
~~Scalar-valued Random Variables~~
Matrix-valued Random Variables



$f(X_1)$ $f(X_2)$ $f(X_3)$ $f(X_4)$

Sample Mean Matrix

$$\frac{1}{4}\left(f(X_1) + f(X_2) + f(X_3) + f(X_4)\right)$$



A Matrix Chernoff Bound for Markov Chains

$$\mathbb{P}\left[\lambda_{\min}\left(\frac{1}{k}\sum_{i=1}^{k} f(X_i)\right) \le -\epsilon\right] \quad \text{and} \quad \mathbb{P}\left[\lambda_{\max}\left(\frac{1}{k}\sum_{i=1}^{k} f(X_i)\right) \ge \epsilon\right]$$

Comparison	Chernoff `52	Tropp`12	GLSS`18	Our Result
X	i.i.d scalars	i.i.d matrices	Stationary random walk on an undirected regular graph with spectral expansion λ	Non-stationary random walk on a regular Markov chain with spectral expansion λ
$f(X)$	X	X	$d \times d$ matrix	$d \times d$ matrix
tail prob.	$\exp(-\Omega(k\epsilon^{-2}))$	$d\exp(-\Omega(k\epsilon^{-2}))$	$d\exp(-\Omega(k(1-\lambda)\epsilon^{-2}))$	$d\exp(-\Omega(k(1-\lambda)\epsilon^{-2}))$

A Matrix Chernoff Bound for Markov Chains

Theorem: Let P be an regular Markov chain with state space $[N]$, stationary distribution π and spectral expansion λ. Let $f\colon [N] \to \mathbb{C}^{d\times d}$ be a matrix-valued function such that

1. $\forall X \in [N], f(X)$ is Hermitian and $\|f(X)\|_2 \leq 1$;
2. $\sum_{X\in[N]} \pi_X f(X) = 0$.

Let $(X_1, X_2, \cdots, X_k)$ denote a k-step random walk on P starting from an initial distribution ϕ. Then for $\epsilon \in (0,1)$:

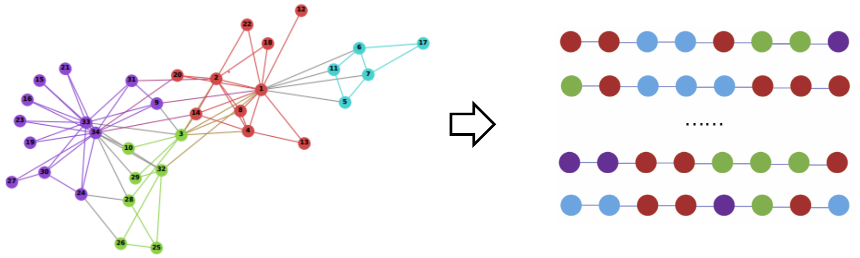
$$\mathbb{P}\left[\lambda_{\min}\left(\frac{1}{k}\sum_{i=1}^{k} f(X_i)\right) \leq -\epsilon\right] \leq \|\phi\|_\pi d^2 \exp(-k(1-\lambda)\epsilon^2/72)$$

$$\mathbb{P}\left[\lambda_{\max}\left(\frac{1}{k}\sum_{i=1}^{k} f(X_i)\right) \geq \epsilon\right] \leq \|\phi\|_\pi d^2 \exp(-k(1-\lambda)\epsilon^2/72)$$

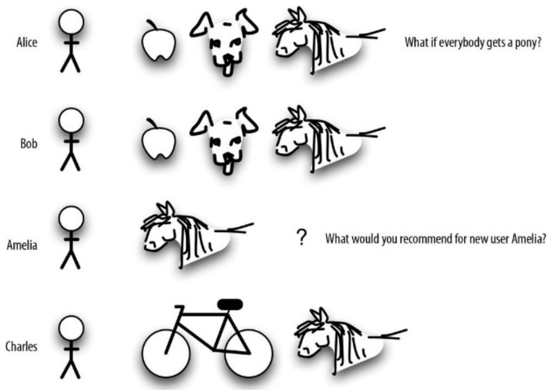
The Application to Co-occurrence Matrices

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

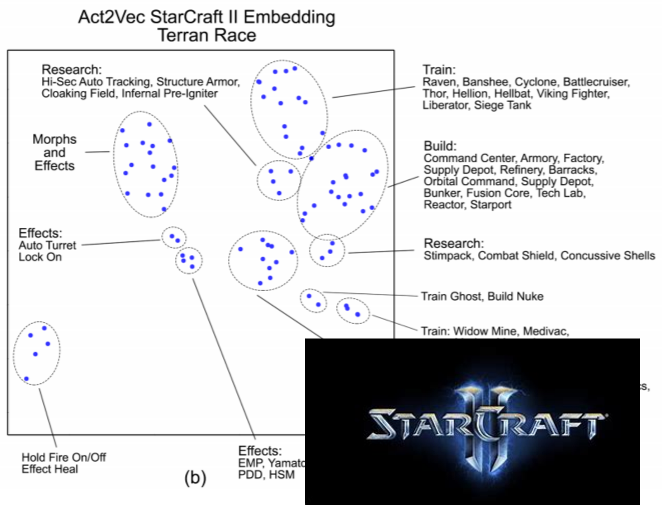
NLP
(LDA, Word2vec, Glove)



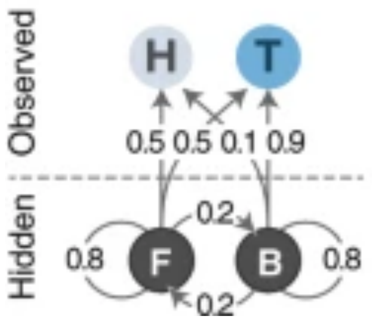
Graph Representation Learning
(DeepWalk, node2vec, metapath2vec)



Recommendation System
(Pin2Vec, Item2vec)



Reinforcement Learning
(Act2Vec)

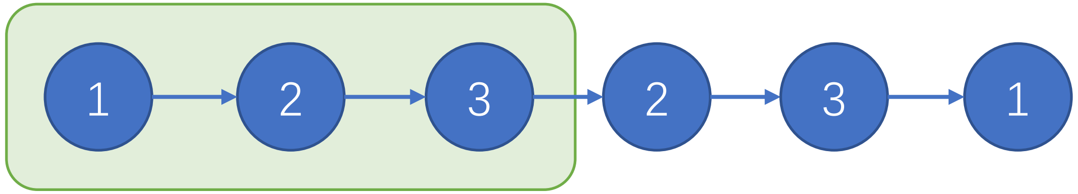


Hidden Markov Models
(Emission Co-occurrence)

Co-occurrence Matrix of Sequential Data

Sliding Window 1

$X_1 = (1,2,3)$

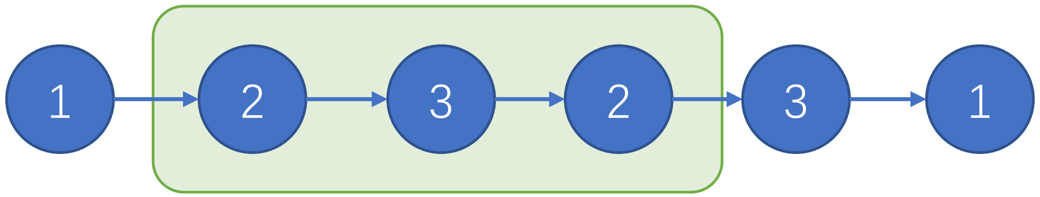


$$\boldsymbol{C} = \frac{1}{4}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Co-occurrence Matrix of Sequential Data

Sliding Window 2

$X_2 = (2,3,2)$

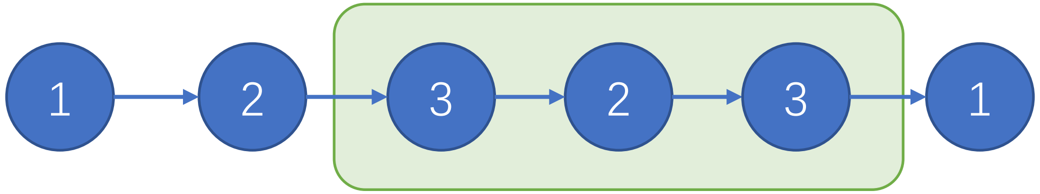


$$\boldsymbol{C} = \frac{1}{2}\left[\frac{1}{4}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix}\right]$$

Co-occurrence Matrix of Sequential Data

Sliding Window 3

$X_3 = (3,2,3)$

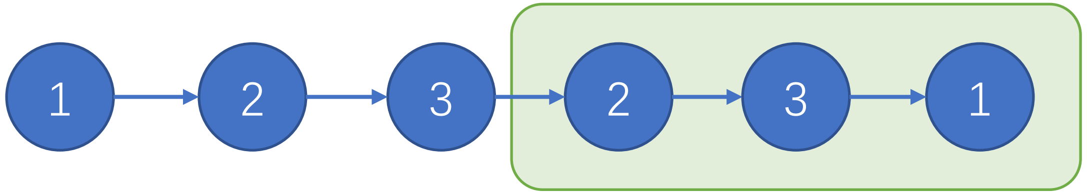


$$\boldsymbol{C} = \frac{1}{3}\left[\frac{1}{4}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\right]$$

Markov chain Matrix Chernoff Bound!

Sliding Window 4

$X_4 = (2,3,1)$



$$\boldsymbol{C} = \frac{1}{4}\left[\frac{1}{4}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}\right]$$

$$= \frac{1}{4}(\boldsymbol{f}(\boldsymbol{X_1}) + \boldsymbol{f}(\boldsymbol{X_2}) + \boldsymbol{f}(\boldsymbol{X_3}) + \boldsymbol{f}(\boldsymbol{X_4}))$$

Observation 1:

Let $\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_{L-T}}$ be the sequence of sliding windows, and $\boldsymbol{f}$ maps a sliding window to the co-occurrence matrix within this window. The co-occurrence matrix C can be written as the **sample mean** of $\boldsymbol{f}(\boldsymbol{X_1}), \boldsymbol{f}(\boldsymbol{X_2}), \cdots, \boldsymbol{f}(\boldsymbol{X_{L-T}})$:

$$\boldsymbol{C} = \frac{1}{L-T}\sum_{k=1}^{L-T} \boldsymbol{f}(\boldsymbol{X_k})$$

Observation 2: If the input sequence $v_1, v_2, \cdots$ is a Markov Chain, then $X_1, X_2, \cdots$ is a Markov Chain, too.

Convergence Rate of Co-occurrence Matrices

- The co-occurrence matrix:

$$C = \frac{1}{L-T}\sum_{k=1}^{L-T} f(X_k)$$

- The asymptotic expectation of C (denote $\Pi = \mathrm{diag}(\pi)$):

$$\mathbb{AE}[C] = \lim_{L\to+\infty} \mathbb{E}[C] = \sum_{r=1}^{T} \frac{1}{2T}\left(\Pi P^r + (\Pi P^r)^\top\right)$$

Theorem: Let P be a regular Markov chain with state space $[n]$, stationary distribution π and mixing time τ. Let $(v_1, \cdots, v_L)$ be a L-step random walk on P starting from a distribution ϕ. Given $\epsilon \in (0,1)$, the probability that the co-occurrence matrix C deviates from its asymptotic expectation $\mathbb{AE}[C]$ (in 2-norm) is bounded by:

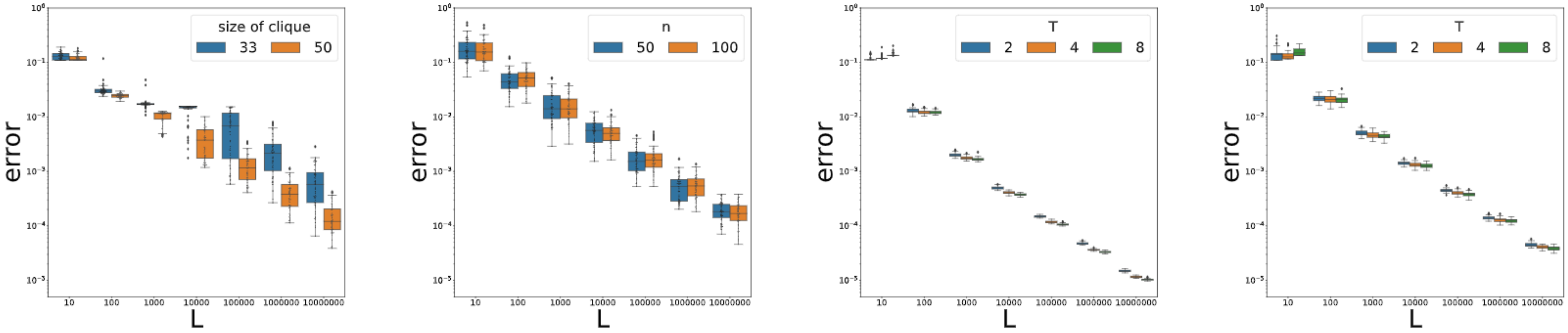
$$\mathbb{P}(\|C - \mathbb{AE}[C]\|_2 \geq \epsilon) \leq 2(\tau + T)\|\phi\|_\pi n^2 \exp\left(-\frac{\epsilon^2(L-T)}{576(\tau+T)}\right)$$

Roughly, one needs $L = O(\tau(\log n + \log\tau)/\epsilon^2)$ samples to guarantee good estimation to the co-occurrence matrix.

Comparison

	X	$f(X)$	Tail Prob.
Chernoff `52	i.i.d scalars	$\boldsymbol{X}$	$\exp(-\Omega(k\epsilon^{-2}))$
Tropp`12	i.i.d matrices	$\boldsymbol{X}$	$d\exp(-\Omega(k\epsilon^{-2}))$
GLSS`18	Stationary random walk on an undirected regular graph with spectral expansion λ	$\boldsymbol{d}\times\boldsymbol{d}$ matrix	$d\exp(-\Omega(k(1-\lambda)\epsilon^{-2}))$
Ours	Non-stationary random walk on a regular Markov chain with spectral expansion λ	$\boldsymbol{d}\times\boldsymbol{d}$ matrix	$d\exp(-\Omega(k(1-\lambda)\epsilon^{-2}))$
HKS`15	Size-1 sliding windows on a reversible Markov chain on $[n]$ with mixing time τ	Co-occurrence matrix within window	$\tau n\exp(-\Omega(L\epsilon^{-2}/\tau))$
Ours	Size-T sliding windows on a regular Markov chain on $[n]$ with mixing time τ	Co-occurrence matrix within window	$(\tau+T)n\exp(-\Omega(L\epsilon^{-2}/(\tau+T)))$

Numerical Experiments



(a) Barbell Graph (b) Winning Streak Chain (c) BlogCatalog (d) Random Graph

Figure 1: The convergence rate of co-occurrence matrices on Barbell graph, winning streak chain, BlogCatalog graph , and random graph (in log-log scale). The x-axis is the trajectory length L and the y-axis is the approximation error $\|\boldsymbol{C} - \mathbb{AE}[\boldsymbol{C}]\|_2$. Each experiment contains 64 trials, and the error bar is presented.

Thanks!

https://arxiv.org/abs/2008.02464