Efficient PageRank Tracking in Evolving Networks

KDD'15

21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining

大坂 直人 (東京大学／プロジェクト RA)
前原 貴憲 (静岡大学)
河原林 健一 (NII)

はじめに

PageRankとPersonalized PageRank

- PageRank [Brin-Page.'98]

  Webページの**重要度**の指標

  リンク構造だけから決まる

一般化

- Personalized PageRank [Jeh-Widom.'03]

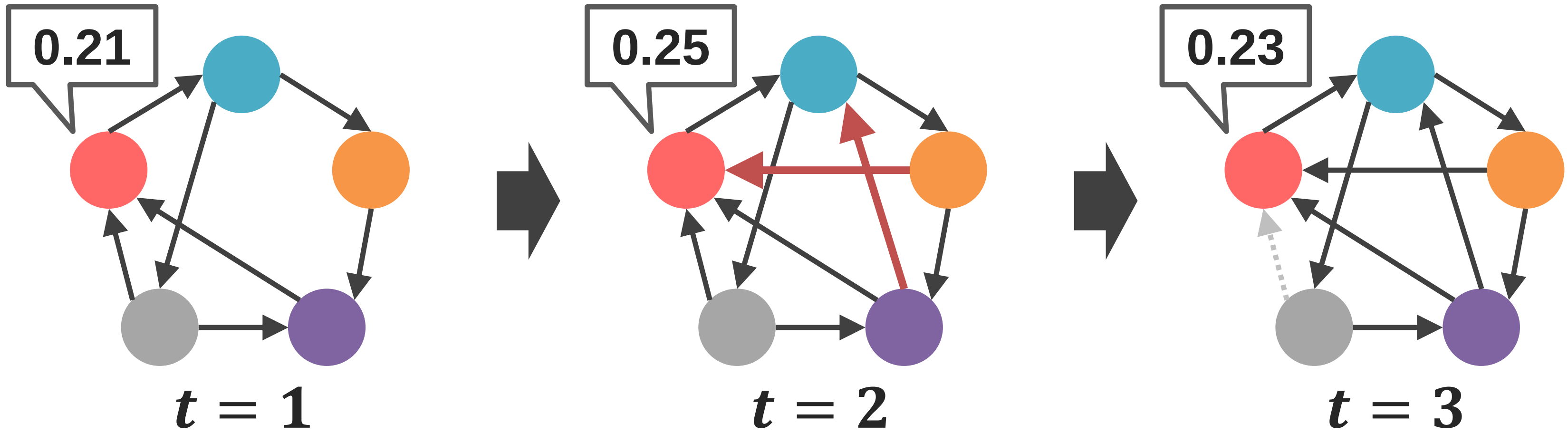
  バイアス付き ⇨ **関連度**

昨年の感謝祭…静的グラフ上の高速計算

[Maehara-Akiba-Iwata-Kawarabayashi. PVLDB'14]

はじめに

Evolving Networks … 動的グラフ

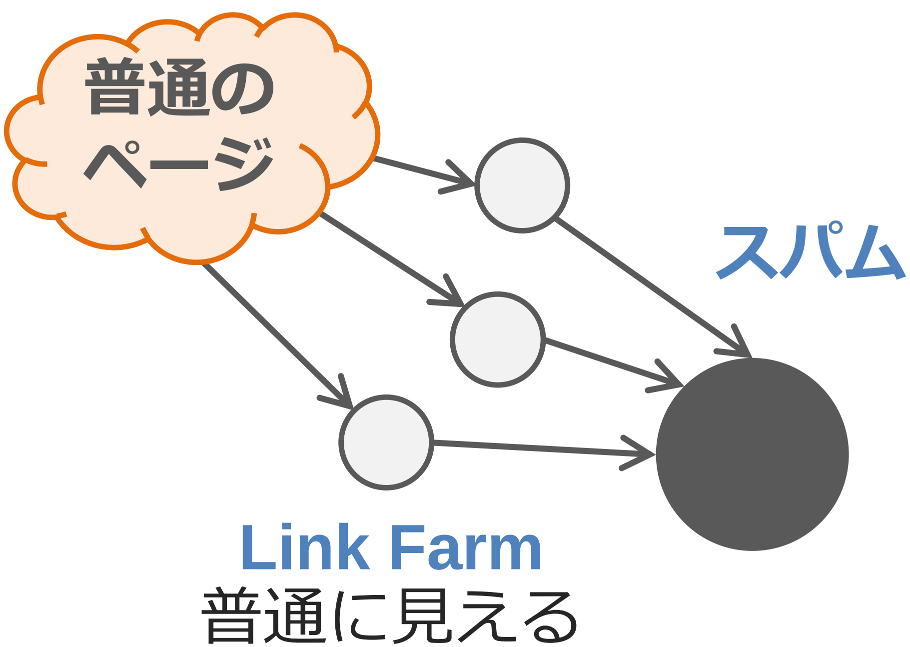


- World Wide Web
  新しいページ・リンク　60万ページ／秒
  http://www.internetlivestats.com/
- ソーシャルネットワーク
  新しい友人関係
- マイクロブログ
  ユーザ同士のやりとり　5000ツイート／秒
  http://www.technologyreview.com/graphiti/522376/the-many-tongues-of-twitter/

グラフ全体を見ずに更新したい

はじめに
関連度としての応用

スパム検知
スコアの時間変化を利用
[Chung-Toyoda-Kitsuregawa. ’09, ’10]



ユーザ推薦
友達の候補生成
[Gupta-Goel-Lin-Sharma-Wang-Zadeh. WWW’13]

はじめに

Personalized PageRank追跡の既存手法

	m辺の無作為挿入の時間	スケーラビリティ 0.1秒以下 / 辺 誤差約10^{-9}
Aggregation/Disaggregation [Chien et al. ’04]	$\mathcal{O}(\boldsymbol{m}\|\boldsymbol{S}\| \log 1/\epsilon)$	68M 辺
Monte-Carlo [Bahmani et al. ’10]	$\mathcal{O}(m + \log m\ /\boldsymbol{\epsilon^2})$	68M 辺
Power method ナイーブな手法	$\mathcal{O}(\boldsymbol{m^2} \log 1/\epsilon)$	11M 辺

本研究の貢献

成長するグラフにおける
Personalized PageRank **追跡**のための
高速 & **高精度** な手法を提案

	m辺の無作為挿入の時間	スケーラビリティ 0.1秒以下 / 辺 誤差約10^{-9}
提案手法	平均 ⇩ 最大出次数 $\mathcal{O}(m + \Delta \log m / \epsilon)$	3,700M 辺
Aggregation/Disaggregation [Chien et al. '04]	$\mathcal{O}(m\|S\| \log 1/\epsilon)$	68M 辺
Monte-Carlo [Bahmani et al. '10]	$\mathcal{O}(m + \log m / \epsilon^2)$	68M 辺
Power method ナイーブな手法	$\mathcal{O}(m^2 \log 1/\epsilon)$	11M 辺

予備知識

Personalized PageRank の定義

[Brin-Page. Comput. Networks ISDN Syst.'98] [Jeh-Widom. WWW'03]

- 線形方程式

  次の解

$$\boldsymbol{x} = \underbrace{\alpha}_{\text{非ジャンプ確率} = 0.85}\underbrace{\boldsymbol{P}}_{\text{確率遷移行列}}\boldsymbol{x} + (1 - \alpha)\overbrace{\boldsymbol{b}}^{\text{バイアス}}$$

Random walk による解釈 ⬇ ⬆ 定常分布

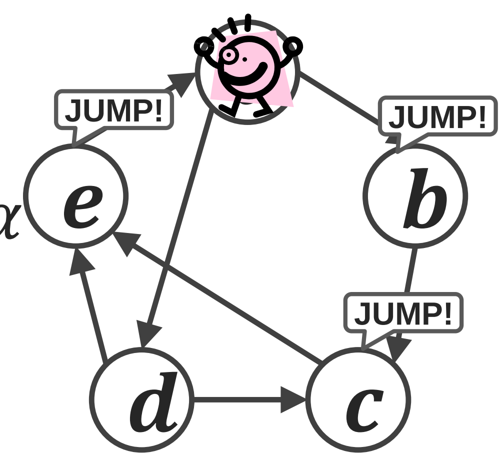
- Random walkによるWeb閲覧のモデル化

  無作為に隣接頂点に移動 *w.p.* α

  無作為に任意頂点にジャンプ *w.p.* $1 - \alpha$

  分布 $b \in \mathbb{R}^n$

  $x_v = v$ を訪れる確率



静的グラフ上のPageRankの計算

- 線形方程式 $x = \alpha P x + (1-\alpha) b$ を解く

  Power method $x^{(\nu)} = \alpha P x^{(\nu-1)} + (1-\alpha) b$

- v を訪れる割合 x_vを見積もる

  Monte-Carlo シミュレーション

動的グラフ上のPageRankの追跡

- Aggregation/disaggregation
  [Chien-Dwork-Kumar-Simon-Sivakumar. Internet Math.’04]
  変化のあった近傍にPower methodを適用
  大きい☹
- Monte-Carlo ベース
  [Bahmani-Chowdhury. VLDB’10]
  Random walkの軌跡を保持・更新
  沢山必要☹

提案手法

提案手法

問題設定

時刻0の**入力**：$G(0)$

$\boldsymbol{G(0)}$

時刻0の**問題**：

$G(0)$ の PageRank $x(0)$ の**近似**計算

$\|\boldsymbol{x}(\boldsymbol{0}) - \boldsymbol{x}^*(\boldsymbol{0})\|_\infty < \boldsymbol{\epsilon}$

⋮

$\boldsymbol{G(t-1)}$

時刻$t-1$の**問題**：

$G(t-1)$ の PageRank $x(t-1)$ の近似計算

時刻tの**入力**：

追加／**削除**される辺集合

$\boldsymbol{G(t)}$

時刻tの**問題**：

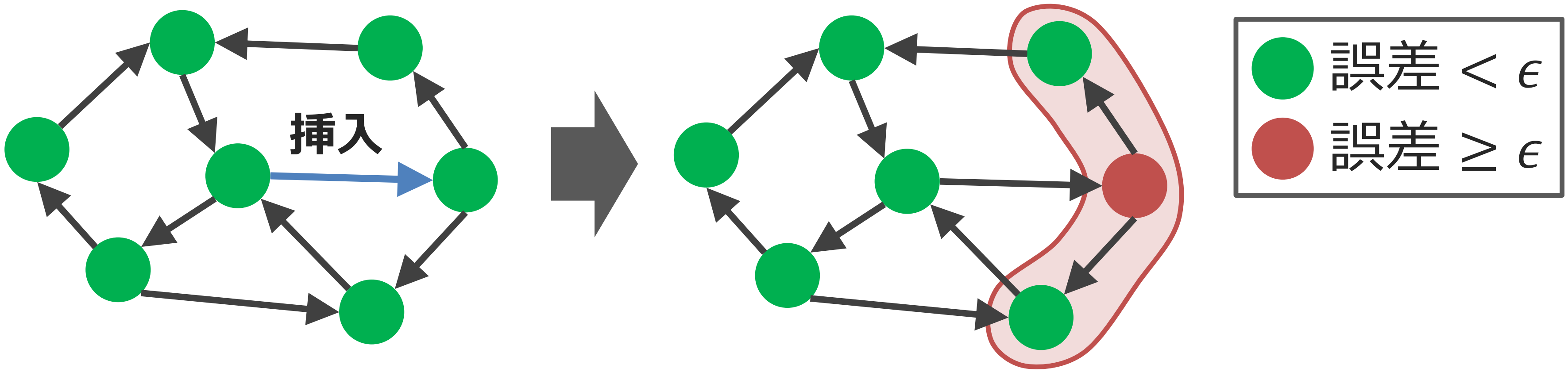
$G(t)$ の PageRank $x(t)$ の近似計算

提案手法

そのアイデア

$$x(t) = \alpha P(t)x(t) + (1-\alpha)b を解く$$

1. $x(t-1)$は$x(t)$の**良い**初期近似解
2. 近似解を**局所的**に改善できないか？



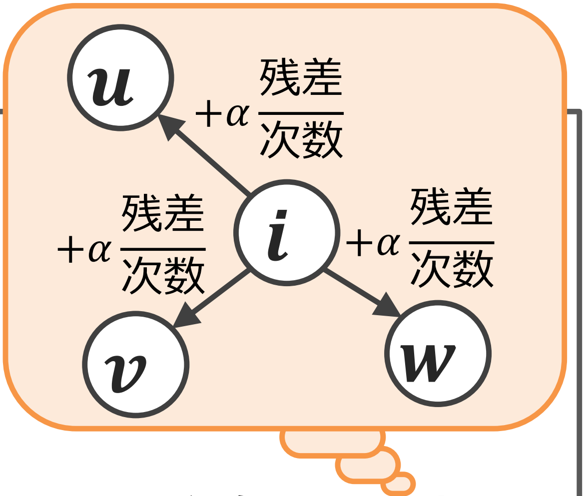
- **Gauss-Southwell** 法 を採用 ☺ [Southwell. ’40, ’46]

  別名 *Local algorithm*
  [Spielman-Teng. SIAM J. Comput.’13]
  Bookmark coloring algorithm
  [Berkhin. Internet Math.’06]

Gauss-Southwell 法 [Southwell. ’40, ’46]

- ν^{th} **近似解** $x^{(\nu)}$
- ν^{th} **残差** $r^{(\nu)} = (1-\alpha)b - (I - \alpha P)x^{(\nu)}$

できるだけ**0**に近づける



$\nu = 1,2,3,\ldots$

$i \leftarrow \left|r_i^{(\nu-1)}\right|$が最大の頂点

If $\left|r_i^{(\nu-1)}\right| < \epsilon$ **terminate**

$r_i^{(\nu)} = 0$ となるよう $x^{(\nu-1)}$ と$r^{(\nu-1)}$を局所的に<u>更新</u>

近似保証: $\left\|\boldsymbol{x}^* - \boldsymbol{x}^{(\nu)}\right\|_\infty \leq \frac{\boldsymbol{\epsilon}}{\boldsymbol{1-\alpha}}$

Gauss-Southwell 法 [Southwell. '40, '46]

- ν^{th} 近似解 $x^{(\nu)}$
- ν^{th} 残差 $r^{(\nu)} = (1-\alpha)b - (I - \alpha P)x^{(\nu)}$

できるだけ$\mathbf{0}$に近づける

$\nu = 1,2,3,\ldots$

$i \leftarrow \left|r_i^{(\nu-1)}\right|$が最大の頂点

If $\left|r_i^{(\nu-1)}\right| < \epsilon$ **terminate**

$x^{(\nu)} = x^{(\nu-1)} + r_i^{(\nu-1)} e_i$

$r^{(\nu)} = r^{(\nu-1)} \boxed{- r_i^{(\nu-1)} e_i + \alpha r_i^{(\nu-1)} P e_i}$

$\leq \dfrac{\|\boldsymbol{r}^{(\mathbf{0})}\|}{(\mathbf{1}-\boldsymbol{\alpha})\boldsymbol{\epsilon}}$ 回

$\boxed{}$ は $\|r^{(\nu-1)}\|_1$ を $(1-\alpha)\epsilon$ 以上減らす

提案手法

その概観

時刻 t:

$$x(t)^{(0)} = x(t-1)$$

$$r(t)^{(0)} = r(t-1) + \alpha\big(P(t) - P(t-1)\big)x(t-1)$$

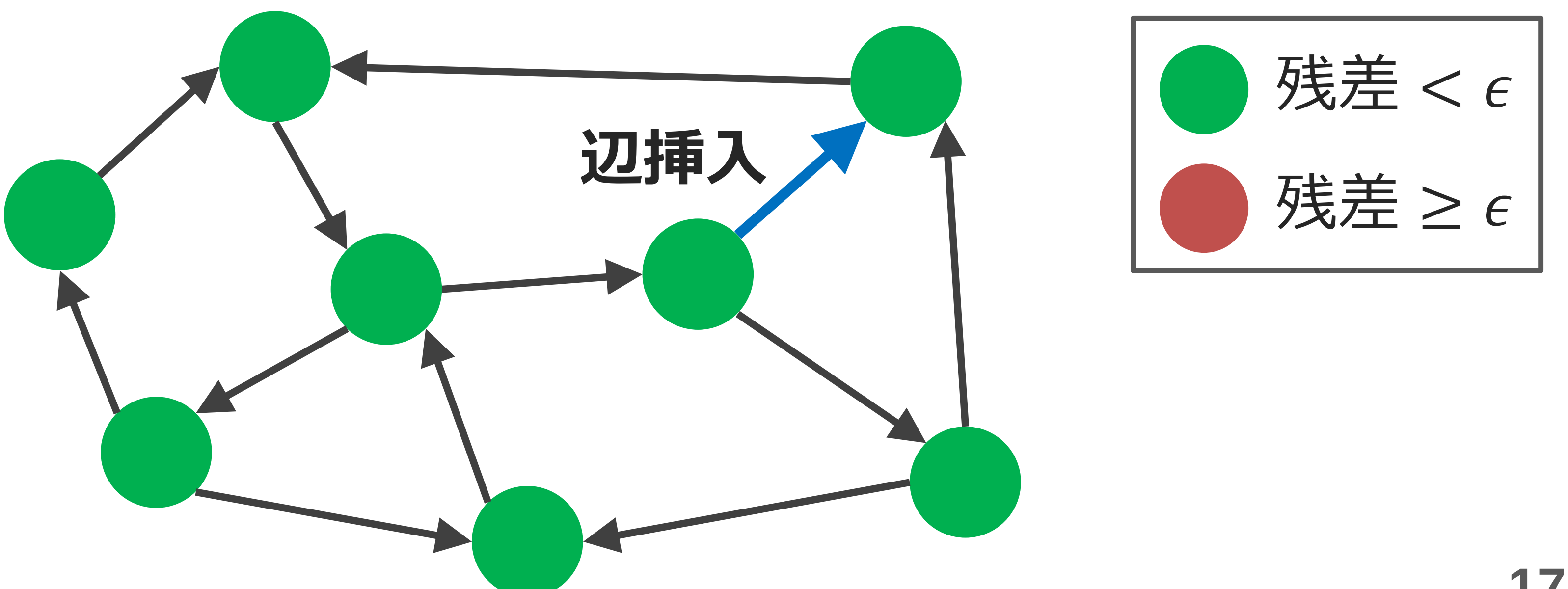
Gauss-Southwell 法を適用

提案手法
挙動

時刻 t:

$$x(t)^{(0)} = x(t-1)$$
$$r(t)^{(0)} = r(t-1) + \alpha\big(P(t) - P(t-1)\big)x(t-1)$$

Gauss-Southwell 法を適用



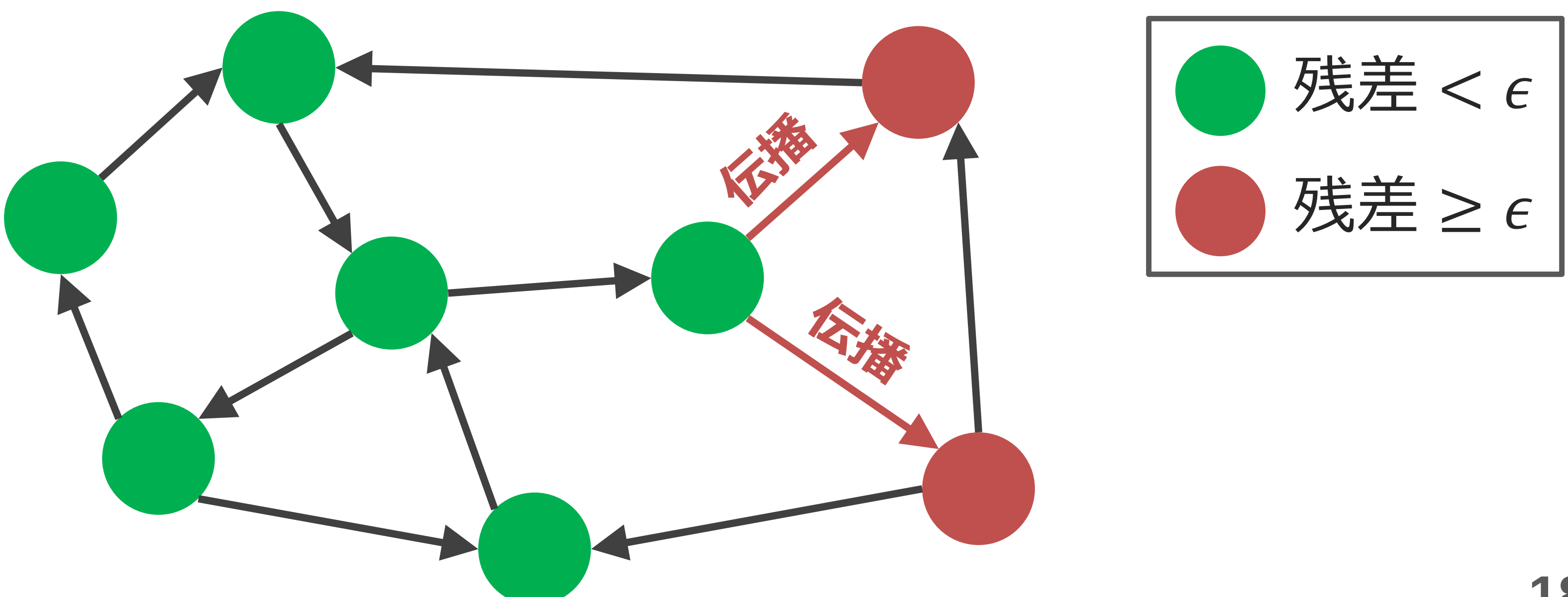
挙動

時刻 t:

$$x(t)^{(0)} = x(t-1)$$

$$\boldsymbol{r}(\boldsymbol{t})^{(0)} = \boldsymbol{r}(\boldsymbol{t}-\mathbf{1}) + \boldsymbol{\alpha}\big(\boldsymbol{P}(\boldsymbol{t}) - \boldsymbol{P}(\boldsymbol{t}-\mathbf{1})\big)\boldsymbol{x}(\boldsymbol{t}-\mathbf{1})$$

Gauss-Southwell 法を適用



提案手法

挙動

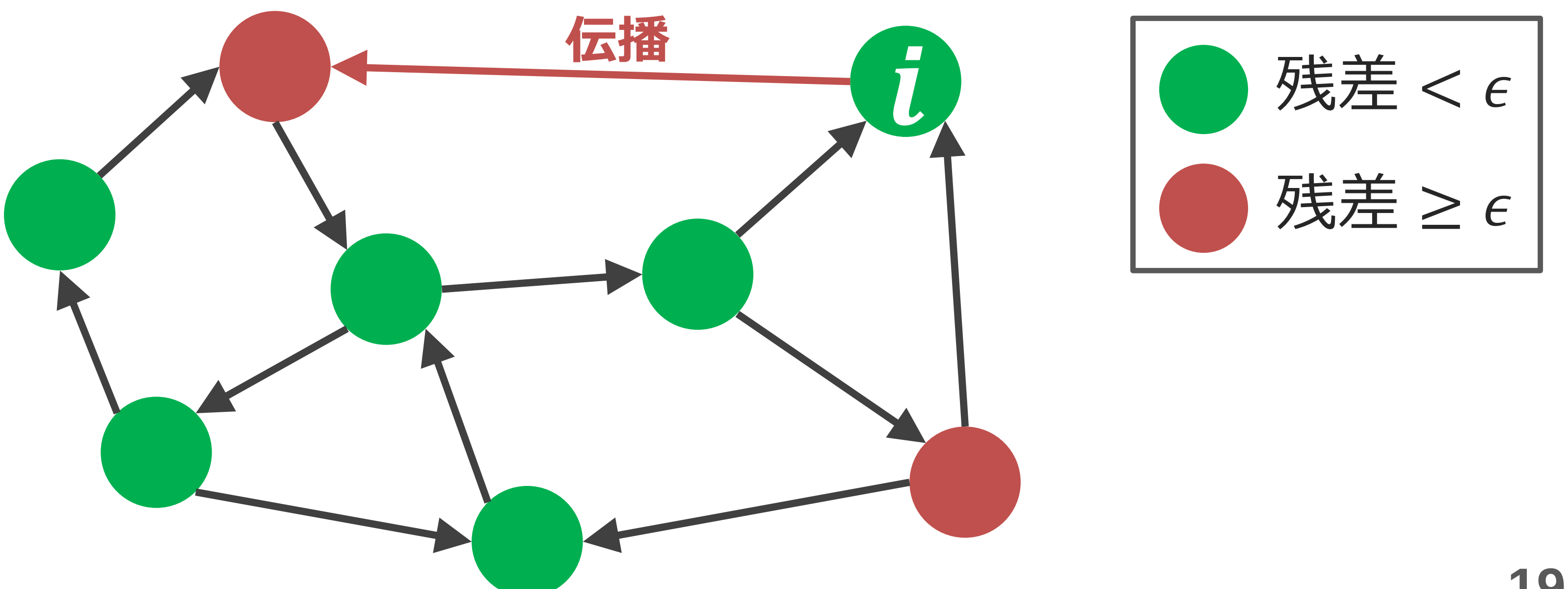
時刻 t:

$$x(t)^{(0)} = x(t-1)$$

$$r(t)^{(0)} = r(t-1) + \alpha\big(P(t) - P(t-1)\big)x(t-1)$$

➡ **Gauss-Southwell 法を適用**

$\boldsymbol{\nu = 1}$



提案手法

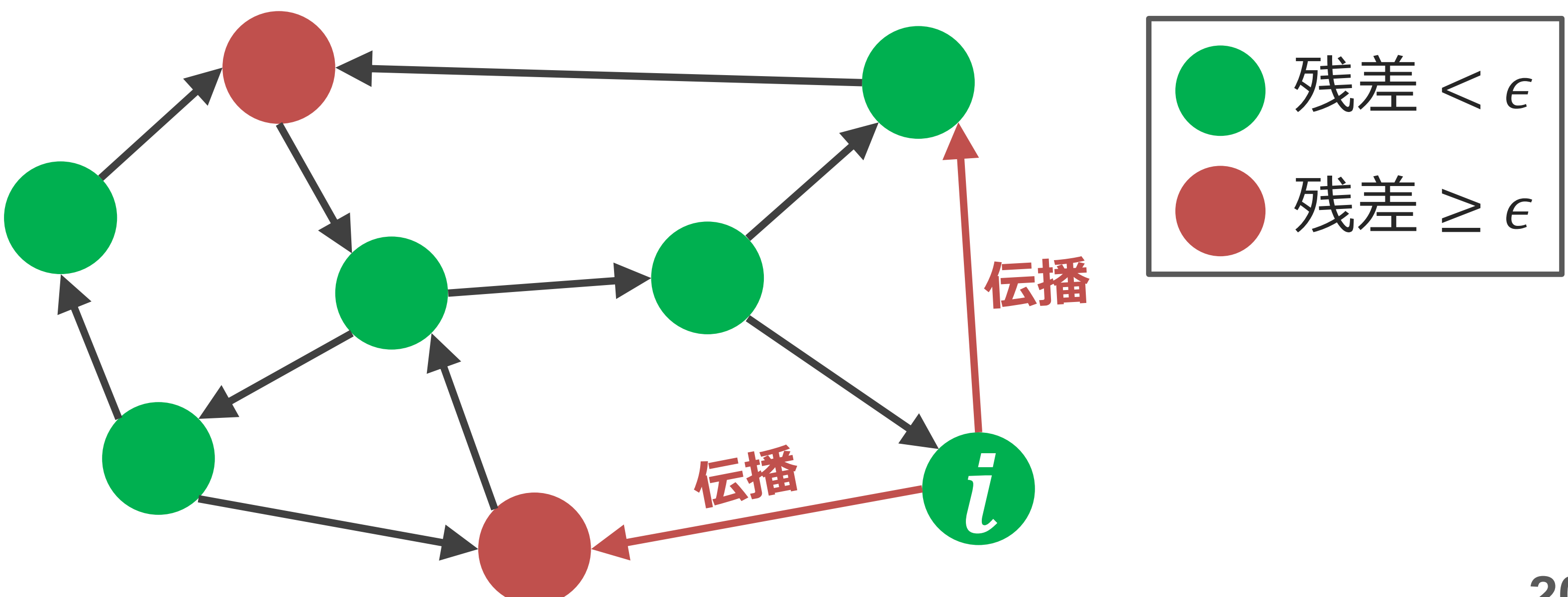
挙動

時刻 t:

$$x(t)^{(0)} = x(t-1)$$
$$r(t)^{(0)} = r(t-1) + \alpha\big(P(t) - P(t-1)\big)x(t-1)$$

➡ **Gauss-Southwell 法を適用**

$\boldsymbol{\nu = 2}$



提案手法

性能解析

時刻 t:

$$x(t)^{(0)} = x(t-1)$$

$$r(t)^{(0)} = r(t-1) + \boxed{\boldsymbol{\alpha}\big(\boldsymbol{P}(\boldsymbol{t}) - \boldsymbol{P}(\boldsymbol{t}-\mathbf{1})\big)\boldsymbol{x}(\boldsymbol{t}-\mathbf{1})}$$

Gauss-Southwell 法を適用

$\boxed{}$ は**効率的**に計算可

辺vw の追加／削除は $\boldsymbol{\mathcal{O}(d_v)}$ 時間

$$\underset{P(t-1)}{\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \mathbf{1/2} & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \mathbf{1/2} & 0 \end{bmatrix}} \Rightarrow \underset{P(t)}{\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1/2 & 0 & 0 & \mathbf{1/3} & 0 \\ 0 & 1 & 0 & \mathbf{1/3} & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \mathbf{1/3} & 0 \end{bmatrix}} \qquad \underset{P(t)-P(t-1)}{\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1/3} & 0 \\ 0 & 0 & 0 & \mathbf{-1/6} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{-1/6} & 0 \end{bmatrix}}$$

性能解析

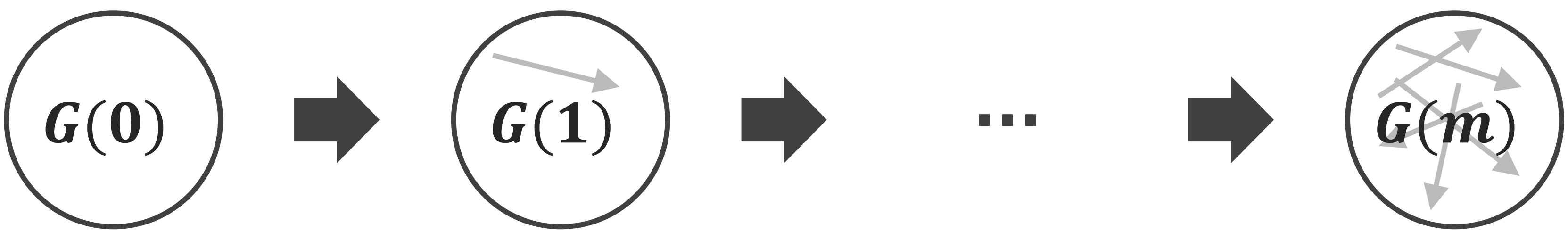
時刻 t:

$$x(t)^{(0)} = x(t-1)$$

$$r(t)^{(0)} = r(t-1) + \overbrace{\alpha\big(P(t) - P(t-1)\big)x(t-1)}^{\textbf{残差の増分}\|\cdot\|_1\textbf{はどの位？}}$$

Gauss-Southwell 法を適用

- 各時刻で単一辺が**無作為**に挿入



辺集合$=\emptyset$　　1辺足す　　m辺足し終えた

$$\mathbf{E}[\text{時刻}t\text{での残差の増分}] \le 2\alpha/t$$

提案手法

性能解析

Gauss-Southwell 法を適用

- **定理 1.**

  任意の変更に対する反復回数は ならし $\mathcal{O}(1/\epsilon)$

  ⇨ ならし時間は $\mathcal{O}(\Delta/\epsilon)$

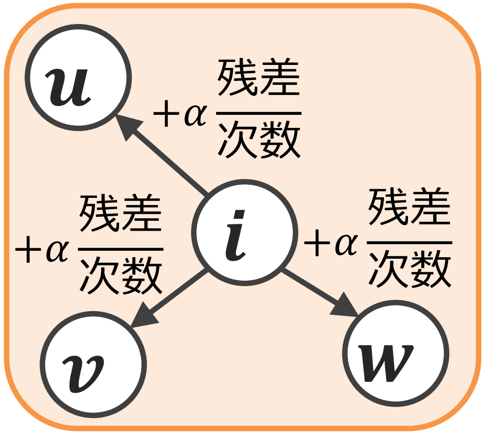
- **定理 2.**

  m辺を**無作為・逐次的**に挿入

  期待総反復回数は $\mathcal{O}(\log m/\epsilon)$

  ⇨ 期待総時間は $\mathcal{O}(m + \Delta \log m/\epsilon)$

Δ = 最大出次数



実験

設定: 単一辺挿入の性能評価

$G(0)$ → $G(1)$ → … → $G(10^5)$

全体のグラフから
100,000辺抜いた

1辺足す

時刻無し: 無作為
時刻付き: 時刻順

100,000辺
足し終えた

- パラメータ設定
  - $\alpha = 0.85$
  - $b = 100$要素が非ゼロのベクトル
  - $\epsilon = 10^{-9}$

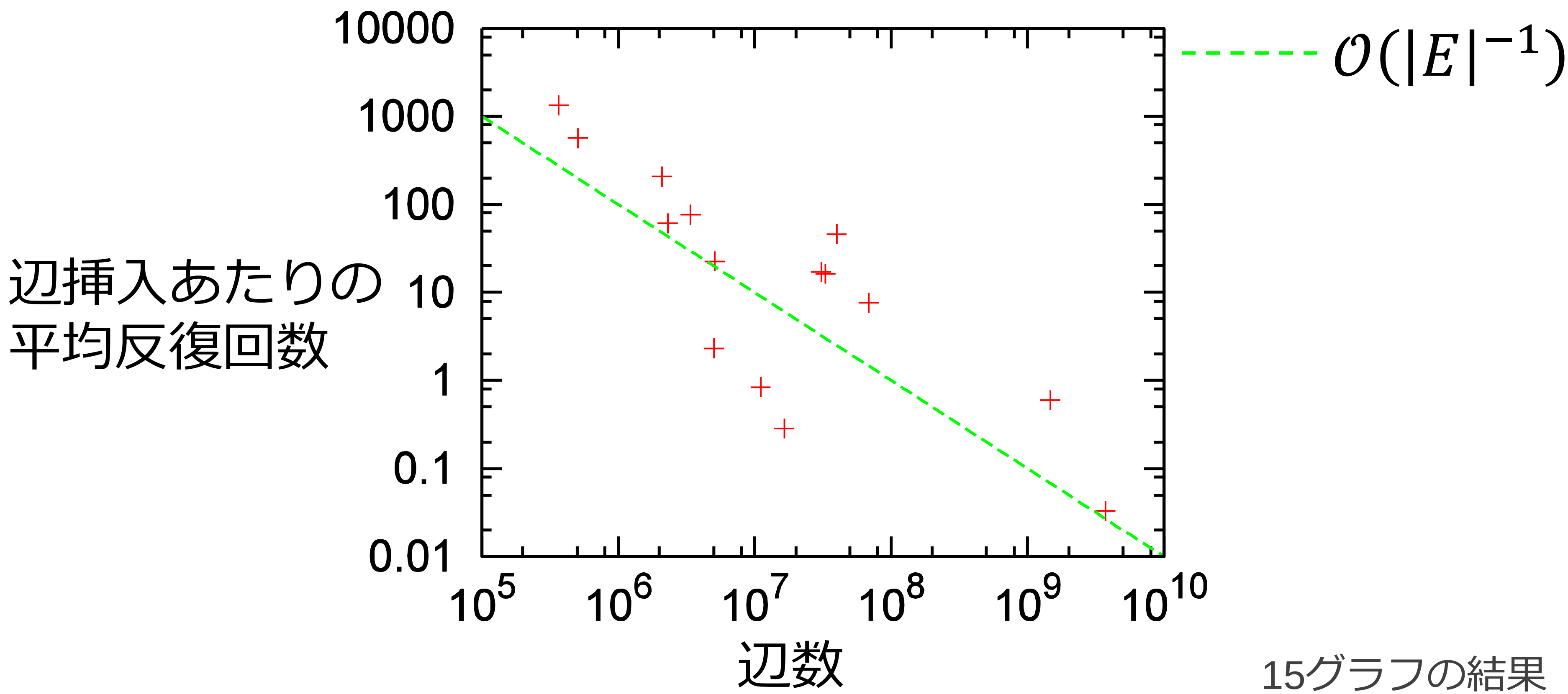
実験
性能評価: 単一辺挿入の平均時間 & 反復回数

データセット [出典]	頂点数 $\lvert V \rvert$	辺数 $\lvert E \rvert$	最大 出次数Δ	平均 時間	平均 反復回数
wiki-Talk [SNAP]	2M	5M	100,022	589.6 μs	2.3
web-Google [SNAP]	1M	5M	3,444	7.2 μs	22.6
as-Skitter [SNAP]	2M	11M	35,387	288.4 μs	0.8
Flickr[時刻] [KONECT]	2M	33M	26,367	95.3 μs	16.2
Wikipedia[時刻] [KONECT]	2M	40M	6,975	76.8 μs	46.0
soc-LiveJournal1 [SNAP]	5M	68M	20,292	17.9 μs	7.6
twitter-2010 [LAW]	42M	1,500M	2,997,469	29,382.8 μs	0.7
uk-2007-05 [LAW]	105M	3,700M	15,402	2.3 μs	0.0

[KONECT] The Koblenz Network Collection http://konect.uni-koblenz.de/networks/
[LAW] Laboratory for Web Algorithmics http://law.di.unimi.it/datasets.php
[SNAP] Stanford Large Network Dataset Collection http://snap.stanford.edu/data/

Environment: Intel Xeon E5-2690 2.90GHz CPU with 256GB memory

性能評価: 辺数と反復回数の関係



15グラフの結果

辺が**多い**ほど少ない

Environment: Intel Xeon E5-2690 2.90GHz CPU with 256GB memory

実験

性能評価: 単一辺挿入の平均時間 & 反復回数

データセット [出典]	頂点数 $\lvert V \rvert$	辺数 $\lvert E \rvert$	最大出次数Δ	平均時間	平均反復回数
wiki-Talk [SNAP]	2M	5M	100,022	589.6 μs	2.3
web-Google [SNAP]	1M	5M	3,444	7.2 μs	22.6
as-Skitter [SNAP]	2M	11M	35,387	288.4 μs	0.8
Flickr[時刻] [KONECT]	2M	33M	26,367	95.3 μs	16.2
Wikipedia[時刻] [KONECT]	2M	40M	6,975	76.8 μs	46.0
soc-LiveJournal1 [SNAP]	5M	68M	20,292	17.9 μs	7.6
twitter-2010 [LAW]	42M	1,500M	2,997,469	29,382.8 μs	0.7
uk-2007-05 [LAW]	105M	3,700M	15,402	2.3 μs	0.0

[KONECT] The Koblenz Network Collection http://konect.uni-koblenz.de/networks/
[LAW] Laboratory for Web Algorithmics http://law.di.unimi.it/datasets.php
[SNAP] Stanford Large Network Dataset Collection http://snap.stanford.edu/data/

Environment: Intel Xeon E5-2690 2.90GHz CPU with 256GB memory

実験

性能評価: 単一辺挿入の平均時間 & 反復回数

データセット [出典]	頂点数 $\lvert V \rvert$	辺数 $\lvert E \rvert$	最大出次数Δ	平均時間	平均反復回数
wiki-Talk [SNAP]	2M	5M	100,022	589.6 μs	2.3
web-Google [SNAP]	1M	5M	3,444	7.2 μs	22.6
as-Skitter [SNAP]	2M	11M	35,387	288.4 μs	0.8
Flickr[時刻] [KONECT]	2M	33M	26,367	95.3 μs	16.2
Wikipedia[時刻] [KONECT]	2M	40M	6,975	76.8 μs	46.0
soc-LiveJournal1 [SNAP]	5M	68M	20,292	17.9 μs	7.6
twitter-2010 [LAW]	42M	1,500M	**2,997,469**	**29,382.8** μs	0.7
uk-2007-05 [LAW]	105M	3,700M	**15,402**	**2.3** μs	0.0

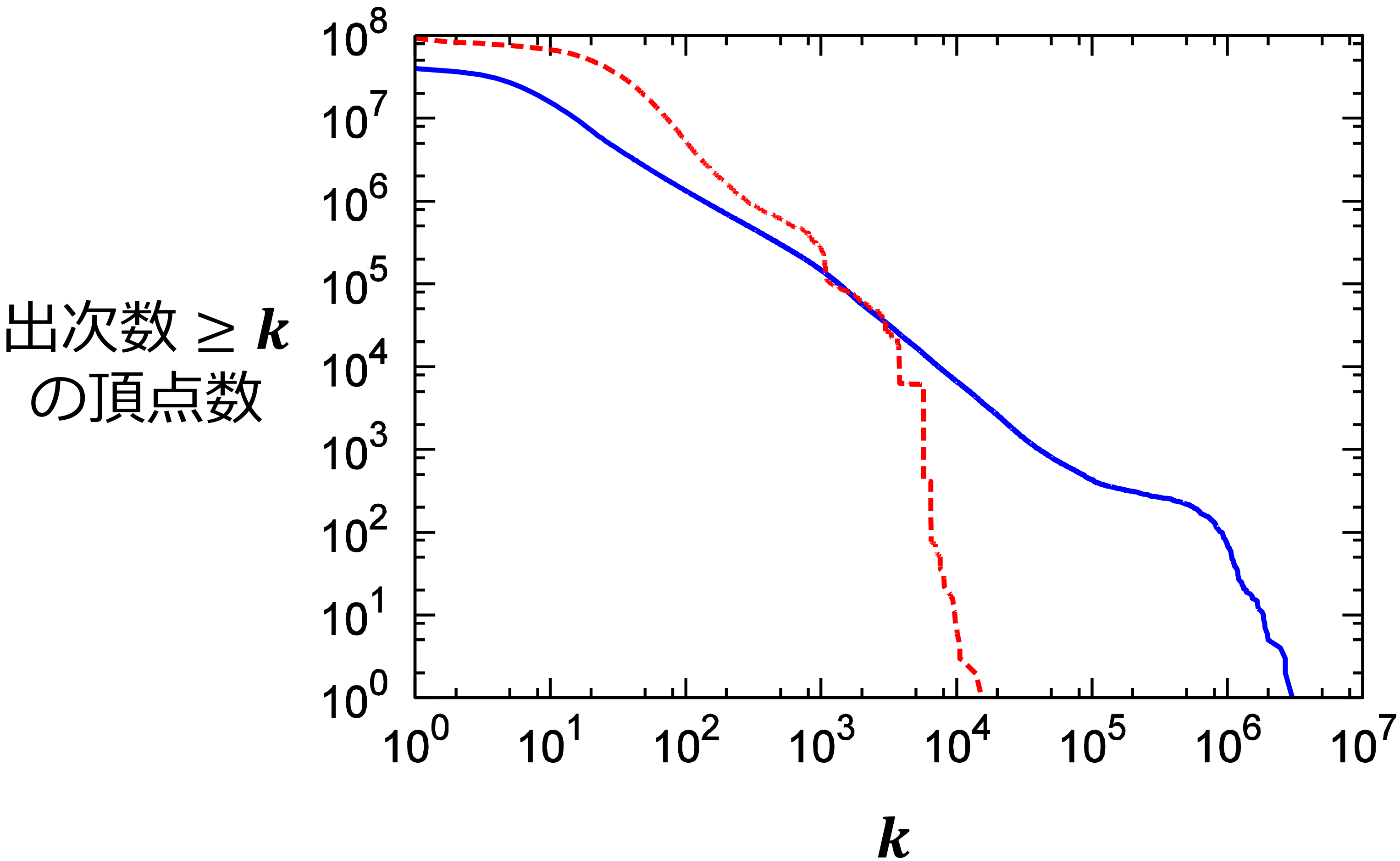
[KONECT] The Koblenz Network Collection http://konect.uni-koblenz.de/networks/
[LAW] Laboratory for Web Algorithmics http://law.di.unimi.it/datasets.php
[SNAP] Stanford Large Network Dataset Collection http://snap.stanford.edu/data/

Environment: Intel Xeon E5-2690 2.90GHz CPU with 256GB memory

次数分布の違い

twitter-2010　(u, v) vがuをフォロー

uk-2007-05　(u, v) ページuからvへリンク



既存手法との比較: 単一辺挿入の平均時間

	web-Google [SNAP] $\lvert V \rvert$=1M $\lvert E \rvert$=**5M**	Wikipedia [KONECT] $\lvert V \rvert$=2M $\lvert E \rvert$=**40M**	uk-2007-05 [LAW] $\lvert V \rvert$=105M $\lvert E \rvert$=**3,700M**
提案手法	**7** μs	**77** μs	**2.3** μs
Aggregation/Disaggregation [Chien et al. '04]	320 μs	40,336 μs	>100,000 μs
Monte-Carlo [Bahmani et al. '10]	444 μs	9,196 μs	>100,000 μs
Warm start power method	80,994 μs	>100,000 μs	>100,000 μs
From scratch power method	>100,000 μs	>100,000 μs	>100,000 μs

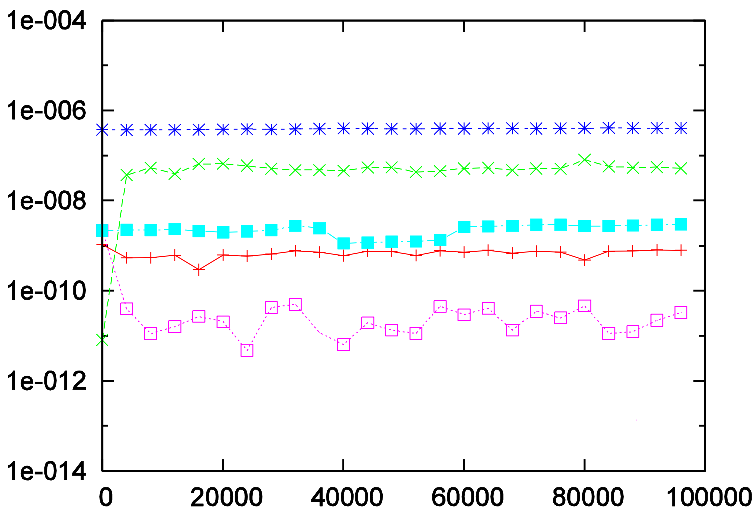
Environment: Intel Xeon E5-2690 2.90GHz CPU with 256GB memory

実験

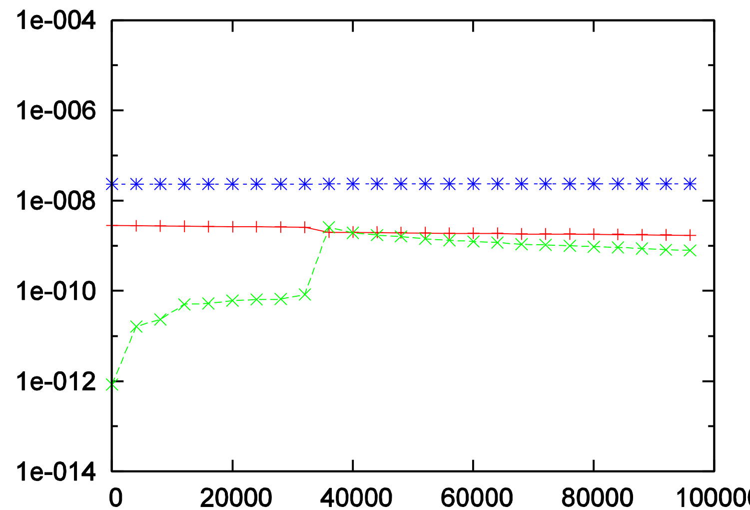
既存手法との比較: 精度

- 提案手法
- Aggregation/Disaggregation [Chien et al.'04]
- Monte-Carlo [Bahmani et al.'10]
- Warm start (power method)
- From scratch (power method)

平均 L_1 誤差の遷移



soc-Epinions1[SNAP] $|V|$=76K $|E|$=509K



Wikipedia[KONECT] $|V|$=2M $|E|$=40M

愚直な手法に匹敵($\sim 10^{-9}$)

Environment: Intel Xeon E5-2690 2.90GHz CPU with 256GB memory

まとめ

成長するネットワークにおける
Personalized PageRank追跡のための
高速 & **高精度**な手法を提案

理論的

- **任意の変更**にならし$\mathcal{O}(\Delta/\epsilon)$時間
- m**辺**の無作為挿入に期待$\mathcal{O}(m + \Delta \log m / \epsilon)$時間
- $\|x - x^*\|_\infty \le \epsilon$

実験的

- **37億辺**をもつグラフへの単一辺挿入に**3μs**
- **10^-9** L_1 誤差

KDD’15は来週シドニー

KDD2015 CALL FOR ATTENDING PROGRAM WORKSHOPS TUTORIALS KDD CUP SPONSORSHIP ORGANISERS

Research Session RT17: Social and Graphs 3

Wednesday 1:00 pm–3:00 pm | Level 3 – Ballroom B

Chair: Tina Eliassi-Rad

Edge-Weighted Personalized PageRank: Breaking A Decade-Old Performance Barrier
Wenlei Xie,Cornell University; David Bindel,Cornell University; Alan Demers,Cornell University; Johannes Gehrke,Cornell University
(Paper ID:117)

SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity
Qingyuan Zhao,Stanford University; Murat A.,Erdogdu; Stanford University Hera,Y.; He Stanford University,Anand; Rajaraman Stanford University,Jure; Leskovec Stanford Universit
(Paper ID:819)

Beyond Triangles: A Distributed Framework for Estimating 3-profiles of Large Graphs
Ethan R.,Elenberg; The University of Texas Karthikeyan,Shanmugam; The University of Texas Michael,Borokhovich; The University of Texas Alexandros,G.; Dimakis The University of Texa
(Paper ID:896)

Scalable Large Near-Clique Detection in Large-Scale Networks via Sampling
Michael Mitzenmacher,Harvard University; Jakub Pachocki,Carnegie Mellon University; Richard Peng,MIT; Charalampos Tsourakakis,Harvard University; Shen Chen Xu,Carnegie Mellon University
(Paper ID:720)

Efficient PageRank Tracking in Evolving Networks
Naoto Ohsaka,The University of Tokyo; Takanori Maehara,Shizuoka University; Ken-ichi Kawarabayashi,National Institute of Informatics
(Paper ID:228)

MASCOT: Memory-efficient and Accurate Sampling for Counting Local Triangles in Graph Streams
Yongsub Lim,KAIST; U Kang,KAIST
(Paper ID:163)