

# 枝刈り探索による ソーシャルネットワークでの 影響最大化アルゴリズム

大坂直人 (東京大学)

秋葉拓哉 (東京大学)

吉田悠一 (NII)

河原林健一 (NII)

EARTO 河原林巨大グラフプロジェクト

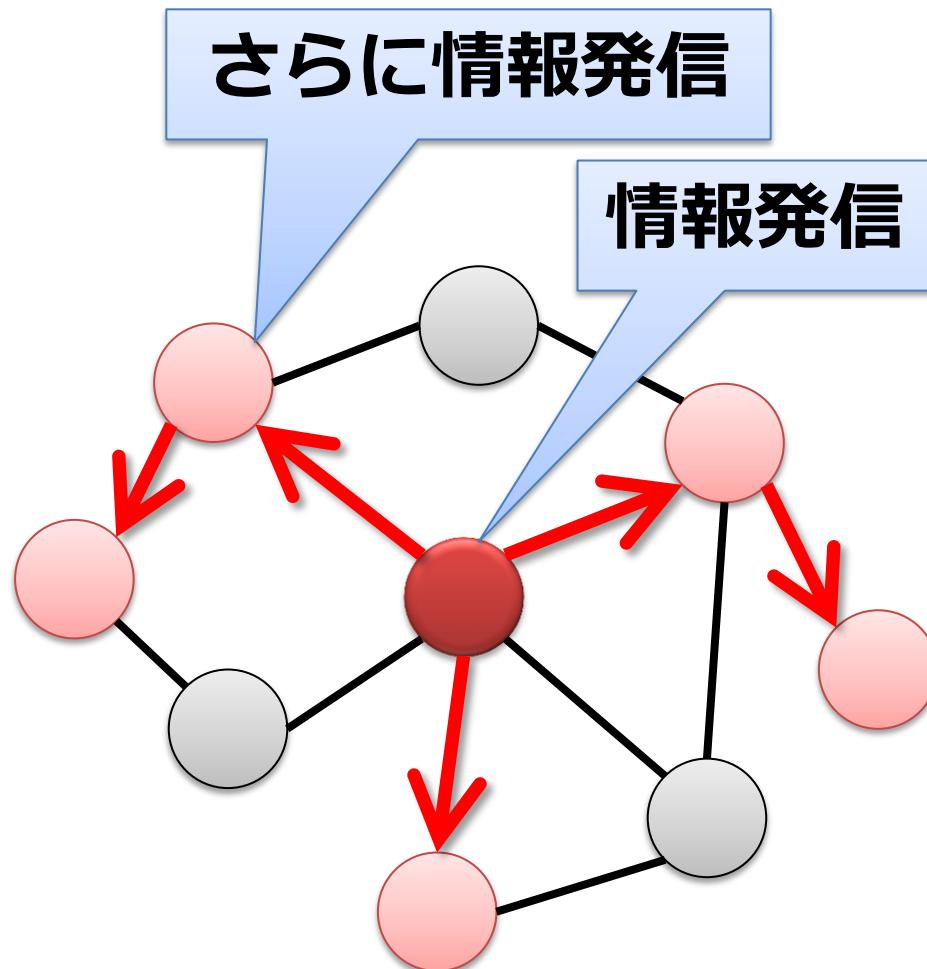
# 情報拡散

ソーシャルネットワーク

- ◆ 頂点…人
- ◆ 辺…関係

- ▶ バイラルマーケティング
- 少数グループに商品の無料サンプルを与え口コミによるプロモーション
- ▶ Twitter上のRT拡散
- ▶ ニュース・世論・噂

...



# バイラルマーケティング

- ▶ 「口コミ」の力を利用したマーケティング
  - ◆ 少数に無料 / 割引商品を提供
  - ◆ 多数に宣伝効果



少ない集団を選び影響力を最大にしたい



## 影響最大化問題

# 独立カスケードモデル上の影響最大化問題の既存手法

	低品質	高品質
低速	シミュレーションベース	貪欲アルゴリズム [Kempe, Kleinberg, Tardos. KDD'03] <b>CELF</b> [Leskovec, Krause, Guestrin, Faloutsos, VanBriesen, Glance. KDD'07] <b>StaticGreedy</b> [Cheng, Shen, Huang, Zhang, Cheng. CIKM'13]
高速	DegreeDiscount [Chen, Wang, Yang. KDD'09] <b>PMIA</b> [Chen, Wang, Wang. KDD'10] <b>SAEDV</b> [Jiang, Song, Cong, Wang, Si, Xie. AAAI'11] <b>IRIE</b> [Jung, Heo, Chen. ICDM'12]	?

ヒューリスティクスベース

# 本研究の貢献

	低品質	高品質
低速	<p>シミュレーションベース 数千万辺のグラフを數十分で処理</p>	<p>貪欲アルゴリズム [Kempe, Kleinberg, Tardos. KDD'03] Faloutsos, Stanoić, Nedevsky [Cheng, Shuai, Huang, Zhang, Cheng. CIKM'13]</p>
高速	<p>DegreeDiscount [Chen, Wang, Yang. KDD'09] PMIA [Chen, Wang, Wang. KDD'10] SAEDV [Jiang, Song, Cong, Wang, Si, Xie. AAAI'11] IRIE [Jung, Heo, Chen. ICDM'12]</p>	<p>提案手法</p>

# 影響最大化問題

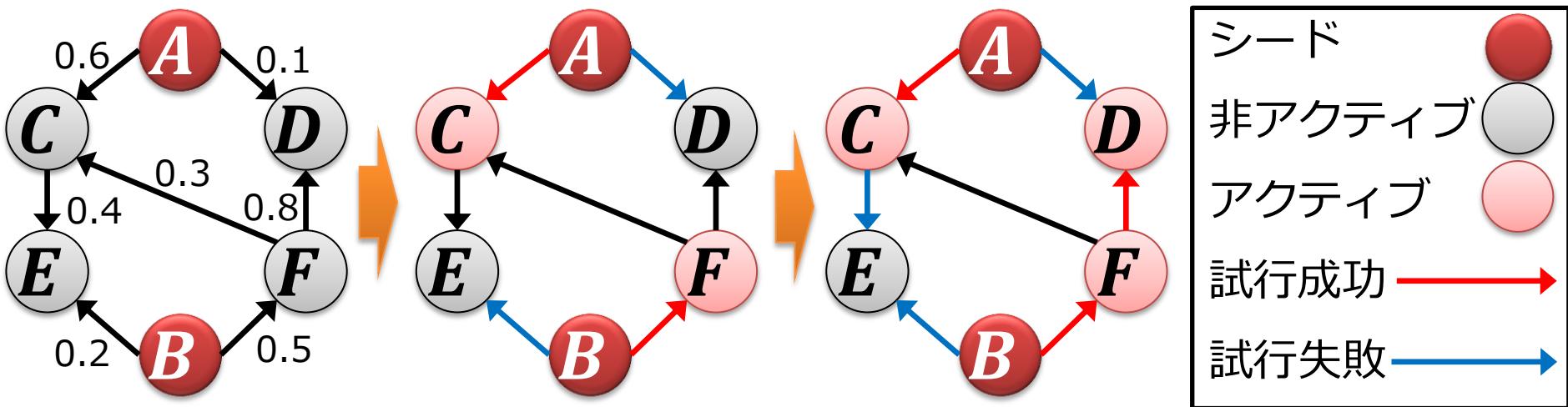
[Kempe, Kleinberg, Tardos. KDD'03]

# 独立カスケードモデル

(Independent Cascade Model)

[Kempe, Kleinberg, Tardos. KDD'03]

入力: グラフ  $G(V, E)$ , 伝播確率  $p: E \rightarrow [0, 1]$ , シード  $S$



0.  $S$ をアクティブに変更
1. アクティブ頂点 $u$ は非アクティブ頂点 $v$ を確率 $p_{uv}$ でアクティブに変更 (試行は1回だけ)
2. 新たなアクティブ頂点がある限り 1. を反復

# 独立カスケードモデル

(Independent Cascade Model)

[Kempe, Kleinberg, Tardos. KDD'03]

入力: グラフ  $G(V, E)$ , 伝播確率  $p: E \rightarrow [0, 1]$ , シード  $S$

$S$  の影響拡散  $\sigma(S) =$   
 $S$  がシードのもと  
アクティブになる頂点数の期待値

$$\sigma(S) = \sum_{\text{outcome } X} \Pr[X] \sigma_X(S)$$

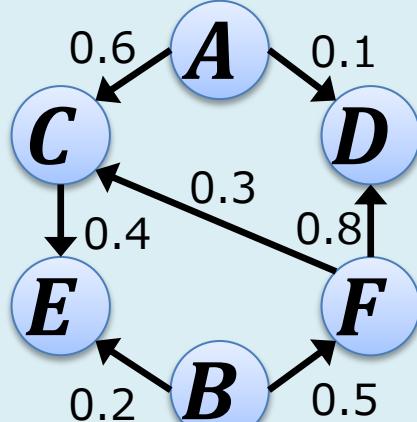
確率  $p_{uv}$  でアクティブに変更 (試行は1回だけ)

2. 新たなアクティブ頂点がある限り 1. を反復

# 影響最大化問題とは

(Influence Maximization Problem)

[Kempe, Kleinberg, Tardos. KDD'03]



$$p_{AC} = 0.6,$$

$$p_{AD} = 0.1,$$

...

$$k = 2$$

グラフ  $G = (V, E)$ , 伝播確率  $p: E \rightarrow [0, 1]$ ,  
シード集合のサイズ  $k$  を入力とし,

$$\max_{S: |S| \leq k} \sigma(S)$$

影響拡散が最大のシード集合を出力する問題

# 既存の結果

## ネガティブ

影響最大化問題は

**NP-hard**

[Kempe, Kleinberg, Tardos. KDD'03]

$\sigma(\cdot)$ の厳密計算は

**#P-hard**

[Chen, Wang, Wang. KDD'10]

解決法

解決法

## ポジティブ

貪欲アルゴリズム

[Kempe, Kleinberg, Tardos. KDD'03]

近似比  $\approx 63\%$

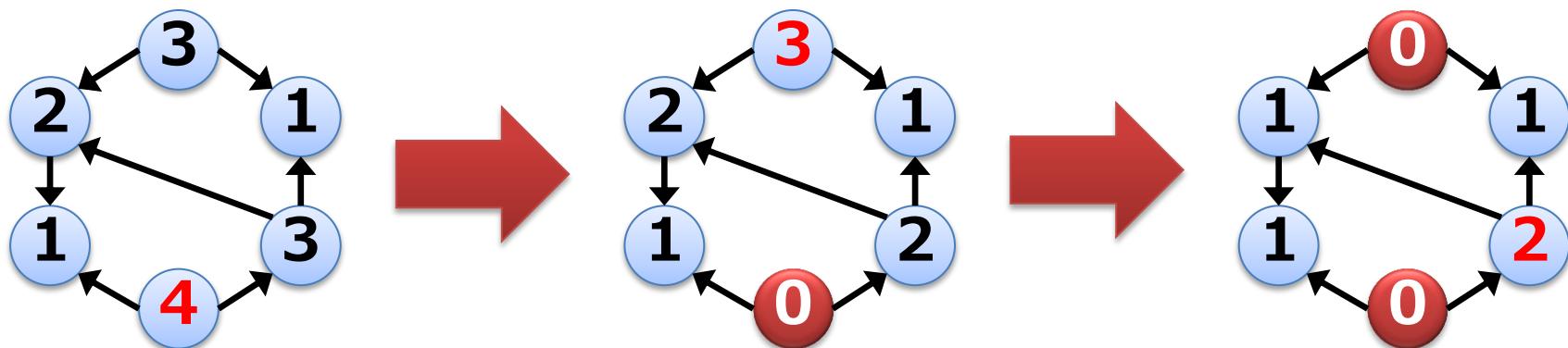
Monte-Carlo  
シミュレーション  
良い近似値を出力

# 貪欲アルゴリズム

[Kempe, Kleinberg, Tardos. KDD'03]

影響拡散の増加量が最大の頂点を  $k$  回選択

$$t \leftarrow \arg \max_{v \in V} \sigma(S \cup \{v\}) - \sigma(S)$$
$$S \leftarrow S \cup \{t\}$$



$$\sigma(S) \geq \left(1 - \frac{1}{e}\right) \text{OPT} \geq 0.63 \times \text{OPT}$$

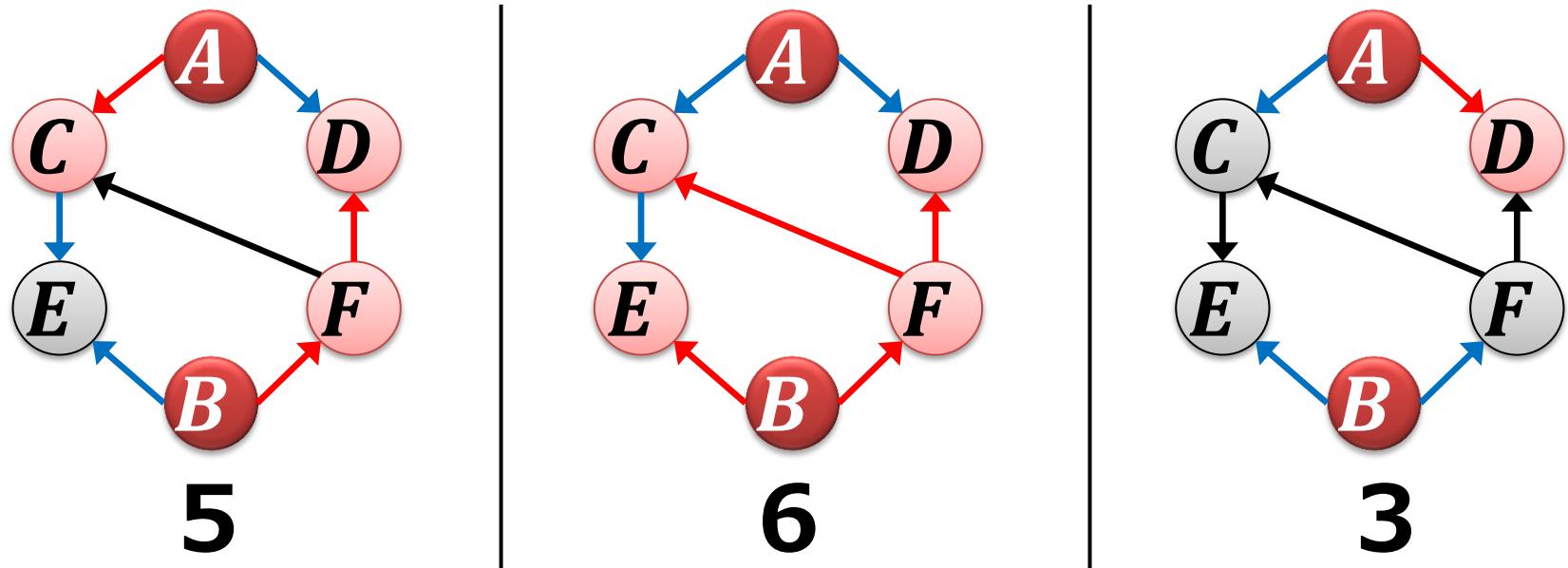
$\sigma(\cdot)$  は劣モジュラ [Kempe, Kleinberg, Tardos. KDD'03]

$\sigma(S \cup \{v\}) - \sigma(S) \geq \sigma(T \cup \{v\}) - \sigma(T) \quad (S \subseteq T \subseteq V, v \in V)$

# Monte-Carloシミュレーション

[Cheng, Shen, Huang, Zhang, Cheng. CIKM'13], [Kempe, Kleinberg, Tardos. KDD'03]  
[Leskovec, Krause, Guestrin, Faloutsos, VanBriesen, Glance. KDD'07]

独立カスケードモデルをシミュレート  
平均アクティブ頂点数を出力



$$\sigma(\{A, B\}) \approx \frac{5 + 6 + 3}{3} \approx 4.67$$

# 問題: 貪欲アルゴリズムはスケーラビリティに乏しい

貪欲アルゴリズム

$\sigma(\cdot)$ の評価回数:  $nk$

Monte-Carlo

シミュレーション

$\sigma(\cdot)$ の計算時間:  $O(mR)$



総計算時間:  $O(knmR)$  ( $R \approx 10,000$ )

$n$ : 頂点数

$m$ : 辺数

$k$ : シードの数

$R$ : シミュレーション回数

数千頂点で限界

# 提案手法

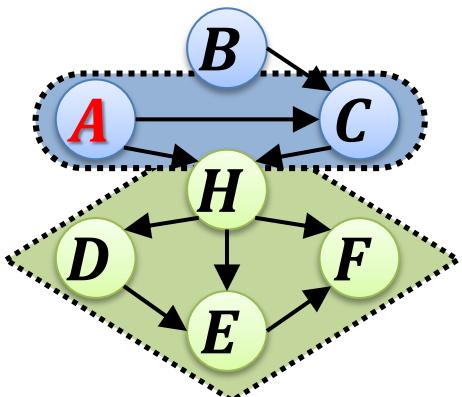
# 概要

## 貪欲アルゴリズム+Monte-Carloシミュレーション

影響拡散の増加量 $\sigma(S \cup \{v\}) - \sigma(S)$ を高速に近似計算

### 枝刈り幅優先探索

各頂点から到達可能な頂点数を効率的に計算



### 「不要な影響拡散の再計算の検知

到達可能な頂点数が不变の頂点を検知  
↓  
不要な幅優先探索を除去

### 「シミュレーション回数の抑制

[Cheng, Shen, Huang, Zhang, Cheng. CIKM'13]によるテクニック

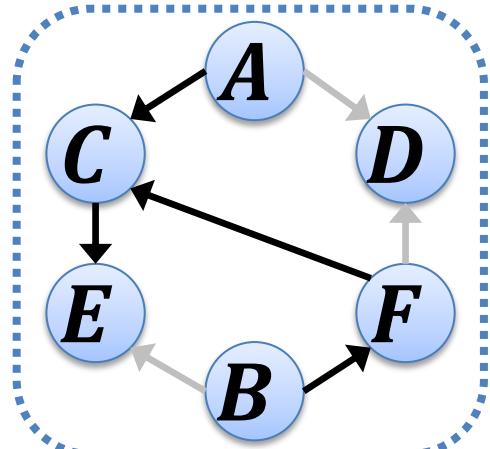
シミュレーション数  
10,000  $\Rightarrow$  200

新たに理論的解析

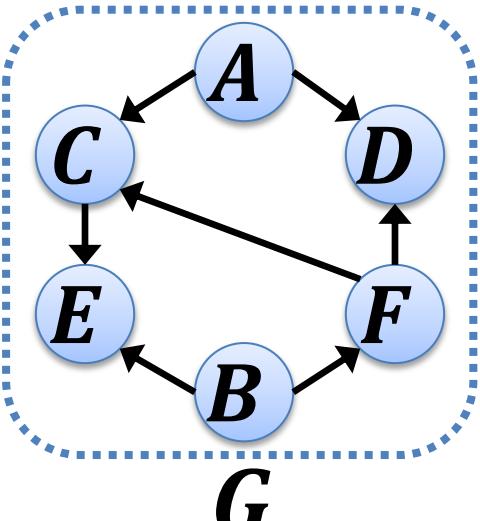
# $\sigma(S)$ の近似: ランダムグラフ生成

各辺 $e$ を確率 $p_e$ で残す

元のグラフ



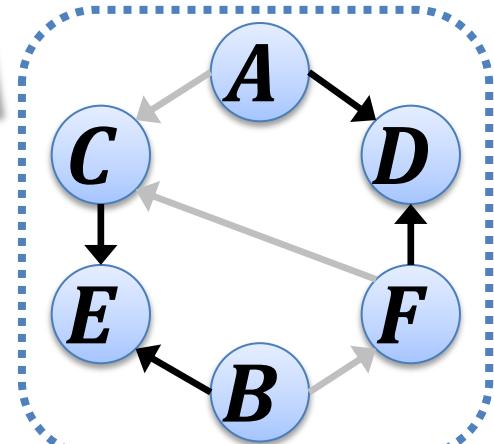
$G_1$



$G$

残った辺: 試行成功  
消えた辺: 試行失敗

$R$  個のランダムグラフ



$G_R$

# $\sigma(S)$ の近似: コインフリップ

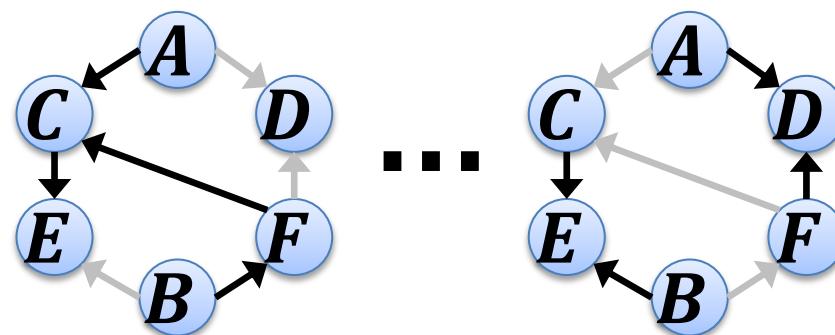
[Kempe, Kleinberg, Tardos. KDD'03]

$$\sigma(S) \approx \frac{1}{R} \sum_{i=1}^R \sigma_{G_i}(S)$$

$\sigma_{G_i}(S)$  =  $G_i$ 上で  $S$  から  
到達可能な頂点数

$v$	$\sigma_{G_1}(\{v\})$	...	$\sigma_{G_R}(\{v\})$	$\sigma(\{v\})$
$A$	3	...	2	2.4
$B$	4	...	2	2.8
$C$	2	...	2	1.6
$D$	1	...	1	1
$E$	1	...	1	1
$F$	3	...	2	2.2

ランダムグラフ数  $R$  大  
↓  
精度 良



# $\sigma(S)$ の近似: コインフリップ

[Kempe, Kleinberg, Tardos. KDD'03]

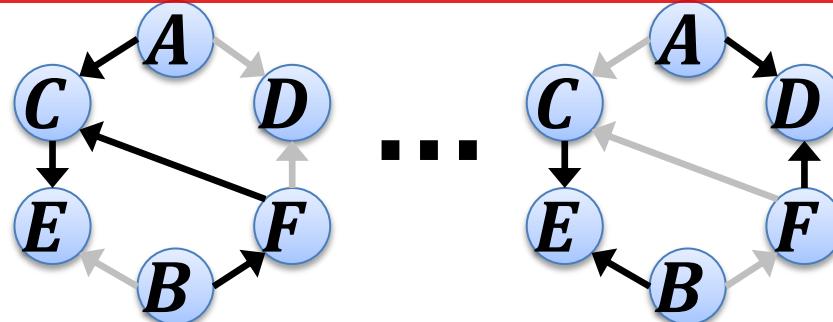
$$\sigma(S) \approx \frac{1}{R} \sum_{i=1}^R \sigma_{G_i}(S)$$

$\sigma_{G_i}(S) =$   $G_i$  上で  $S$  から  
到達可能な頂点数

各頂点から到達可能な  
頂点数を高速に求めたい！

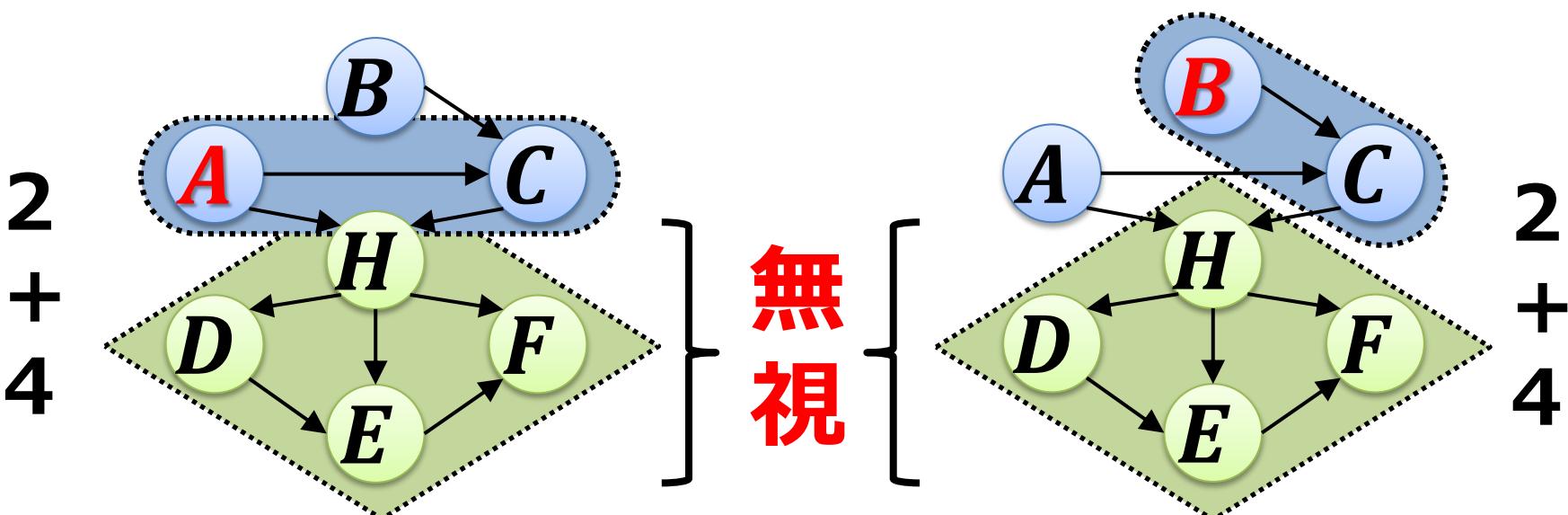
$v$	$\sigma_{G_1}(\{v\})$	...	$\sigma_{G_R}(\{v\})$	$\sigma(\{v\})$
$A$	3	...	2	2.4
$B$	4	...	2	2.8
$C$	2	...	2	1.6
$F$	3	...	1	1
				1
				2.2

ランダムグラフ数  $R$  大  
↓  
精度 良

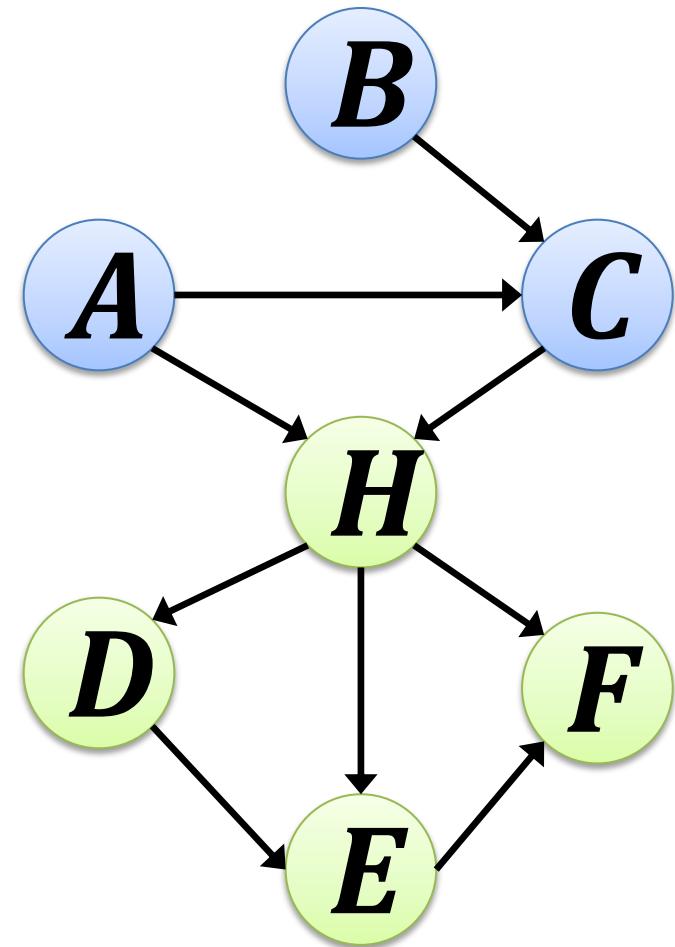


# ハブを利用した枝刈り幅優先探索

- ▶ 前処理
  - ◆ 次数最大の頂点  $H$  の **先祖**と**子孫**を計算
- ▶ 枝刈り（開始頂点  $v$ ）
  - ◆ もし、  $v$  が  $H$  の **先祖**なら、  $H$  の **子孫**を刈る

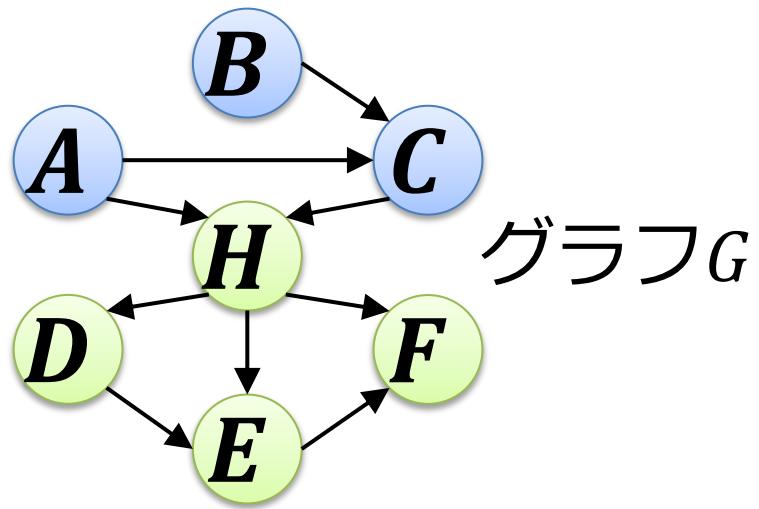


# 枝刈りの効果

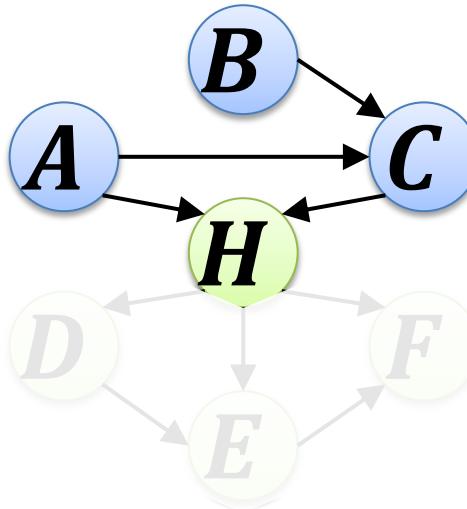


$v$	普通	枝刈り
A	6	2
B	6	2
C	5	1
D	3	3
E	2	2
F	1	1
H	4	4
合計	27	15

# 影響拡散の増加量の計算方法



グラフG



グラフH  
Dから到達可能な  
頂点を除去

同じ  
C

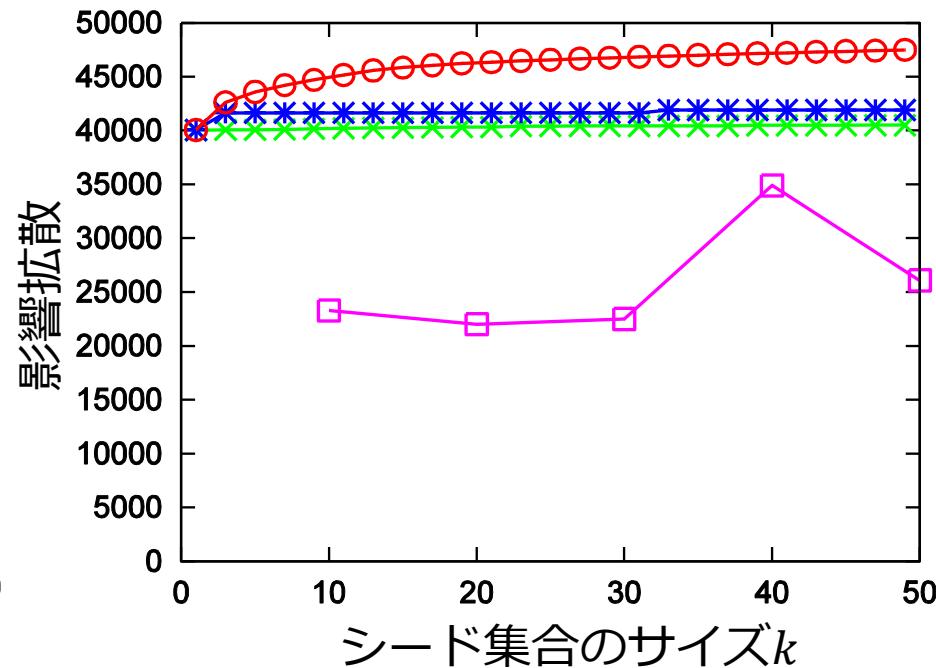
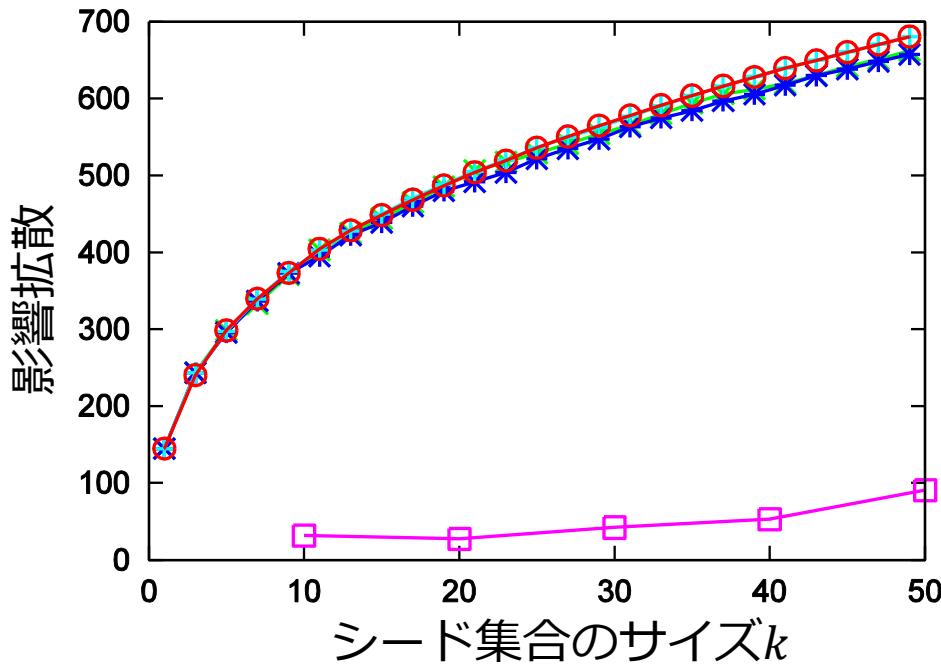
$v$	A	B	C	D	E	F	H
$\sigma_G(\{v\})$	6	6	5	3	2	1	4
$\sigma_G(\{D, v\}) - \sigma_G(\{D\})$	3	3	2	0	0	0	1
$\sigma_H(\{v\})$	3	3	2	0	0	0	1

# 実験結果

# 実験結果: 影響拡散

全ての辺の伝搬確率を  $p_e = 0.01$  に設定

提案手法 —○—  
StaticGreedy(200) —+—  
IRIE —\*—  
PMIA —×—  
SAEDV —□—



データセット	$ V $	$ E $
Epinions(左)	76K	509K
Live Journal(右)	4.8M	69M

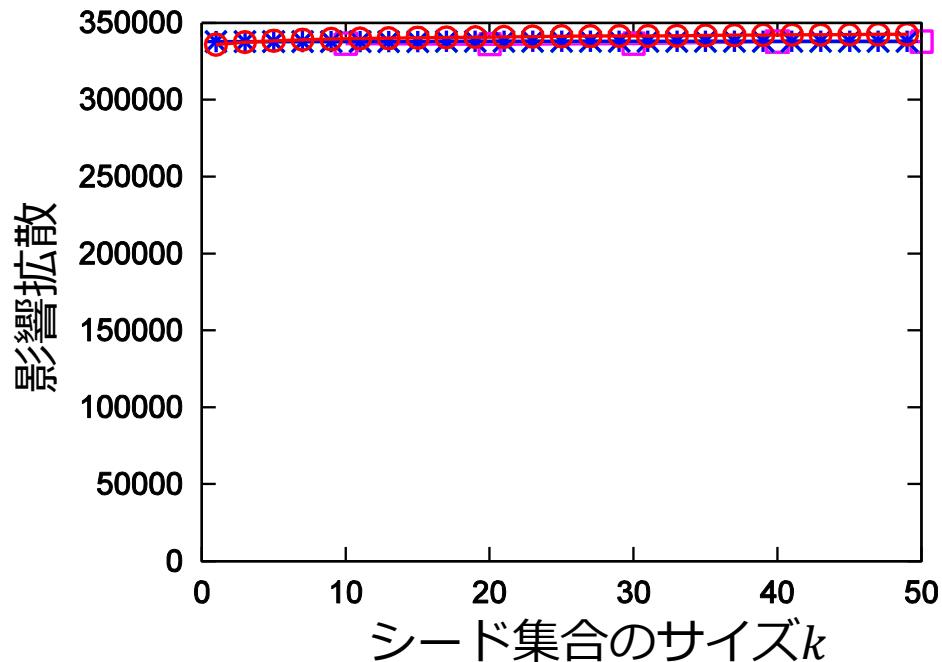
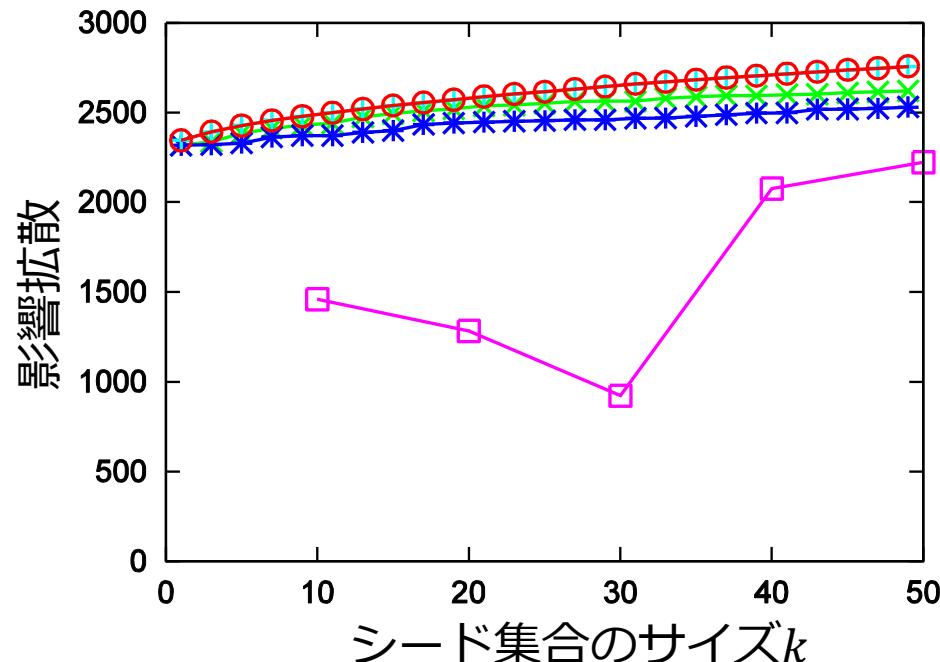
# 実験結果: 影響拡散

全ての辺の伝搬確率を  $p_e = 0.025$  に設定

提案手法  
StaticGreedy(200)

IRIE  
PMIA

SAEDV

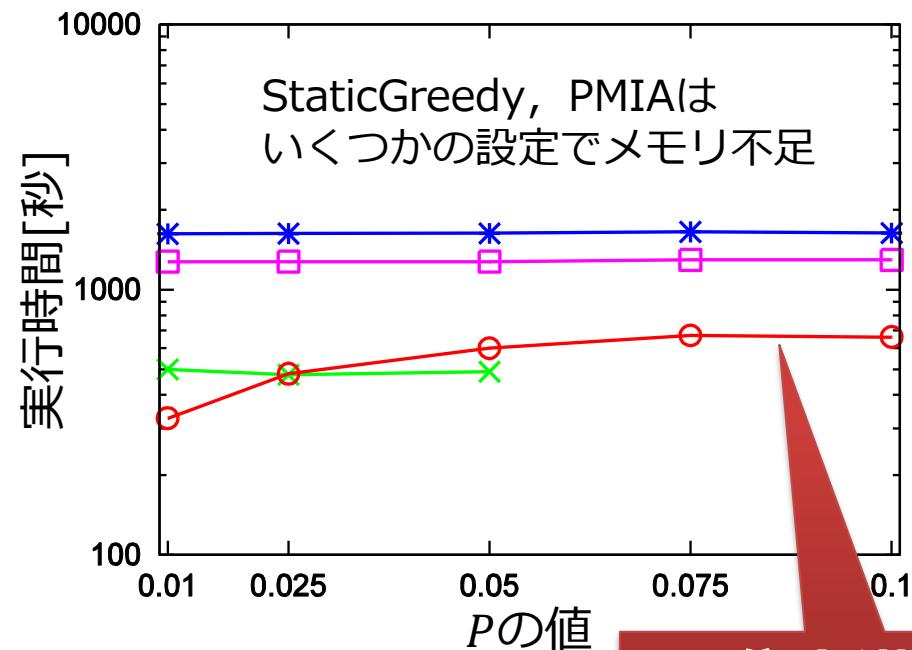
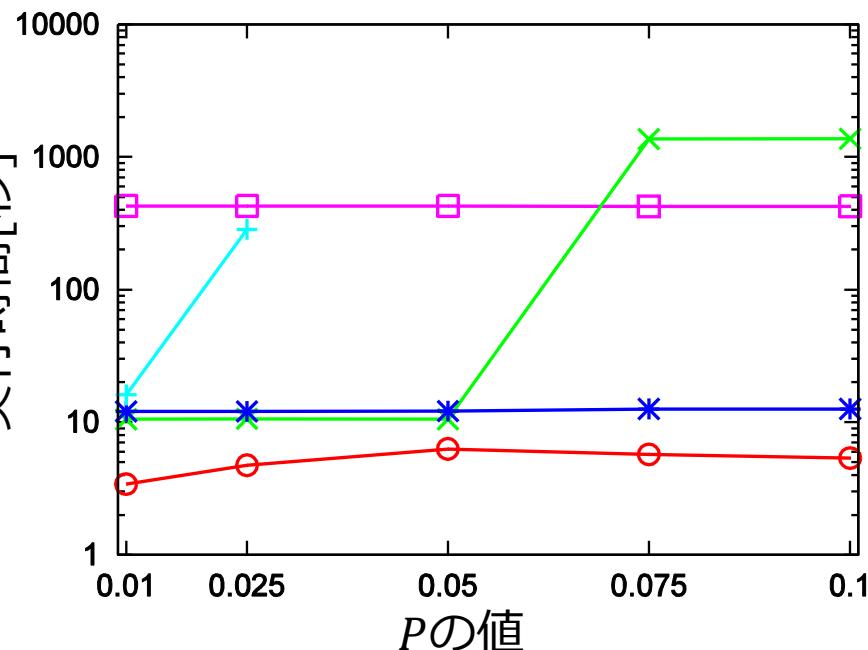


データセット	頂点数	辺数
Epinions(左)	76K	509K
Live Journal(右)	4.8M	69M

# 実験結果: 実行時間

全ての辺の伝搬確率を  $p_e = P$  に設定, シード集合のサイズ50

提案手法 StaticGreedy(200) PMIA IRIE SAEDV



20分未満

データセット	頂点数	辺数
Epinions(左)	76K	509K
Live Journal(右)	4.8M	69M

# まとめ

- ▶ 影響最大化問題の難しさ
  - ◆ 目的関数の 計算量が莫大  
評価回数が多い
- ▶ 影響最大化問題の高速手法を提案
  - ◆ 到達可能性判定に枝刈り幅優先探索を導入
  - ◆ 数十分で数千万辺のネットワークを処理
- ▶ 今後の課題
  - ◆ 並列化…シミュレーションベースの利点
  - ◆ 枝刈り幅優先探索の理論的解析