

令和五年度
高専生インターンシップ 成果報告

発表者 東京都立産業技術高等専門学校 専攻科

河戸幾利

指導教員 北陸先端科学技術大学院大学

林幸雄 教授

研究テーマ1

権力者の中心性を活用したインフルエンサーの抽出

参考にした論文

[1]Fuxuan Liao, Yukio Hayashi.

Applied Network Science volume 7, Article number: 52 (2022)

ネットワークで考える拡散現象

社会を構成する様々なものがネットワーク

- ・ 交通網, 鉄道網, 航空網, WWW, 通信網, 送電網
- ・ 交際関係の輪, 俳優の共演関係, 論文の共著者関係, 製品開発
- ・ 生物の新陳代謝, 脳の神経回路 など...

昨今の課題：感染症の流行

- ・ 現代では航空や鉄道など, 人の移動が長距離かつ激しく行われる
- ・ より多くの人と接触する人もいれば, ほとんど関わらない人もいる
→罹患しているか, 免疫をもっているかの単なる問題ではなく,

疫病の拡散現象はネットワークで考える必要がある

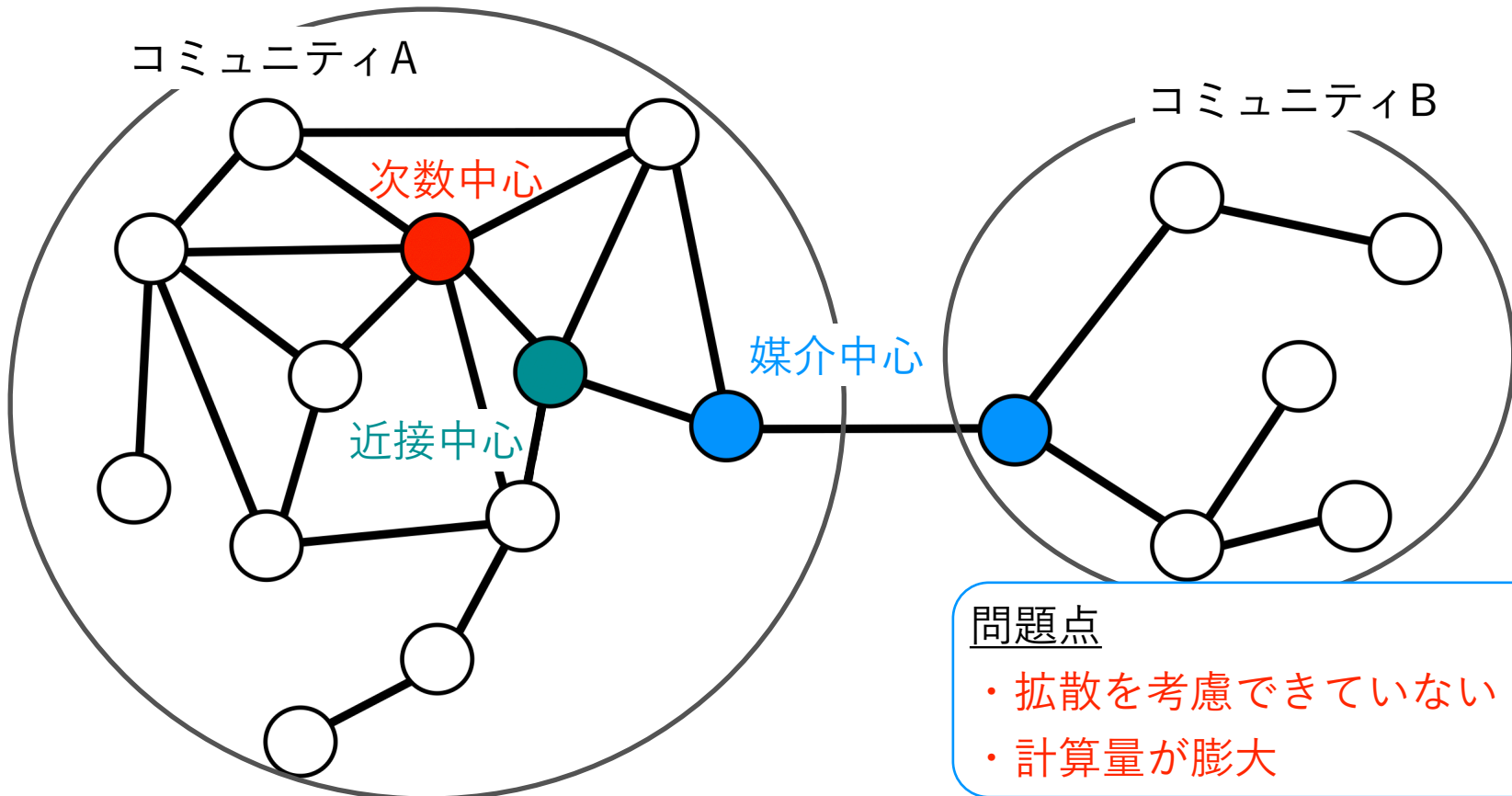
本研究の目的

ネットワークの中から**インフルエンサー**（**拡散影響力の高い中心的なノード**）
を定量的に見つける

中心的なノードを見つける手法

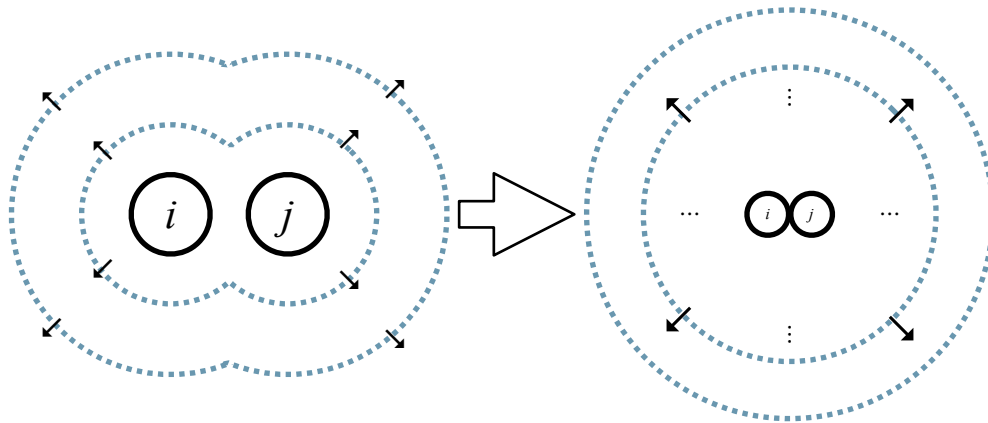
ノードに中心性の指標を与える既存手法

- ・ 次数中心性（ノードの次数の多さ）
- ・ 媒介中心性（経由される頻繁さ）
- ・ 近接中心性（ネットワークの重心） など



距離が離れたシードを選ぶ必要

従来手法では、ノード間の距離を考慮できていない



シードを2つ選んでいるが、ノード間の距離が近いため、拡散が広がると1つとして選んでいることと同義

→**拡散現象を考慮できていない**

先行研究[1]ではMP-MVC:Message Passing-Minimum Vertex Coverが提案

- ・シードの選び方：メッセージ伝搬法による最小頂点被覆
- ・距離が離れたシードを選択：**l-ホップ被覆**
→既存手法よりも優れていることが示されている

この手法を参考として本研究を行った

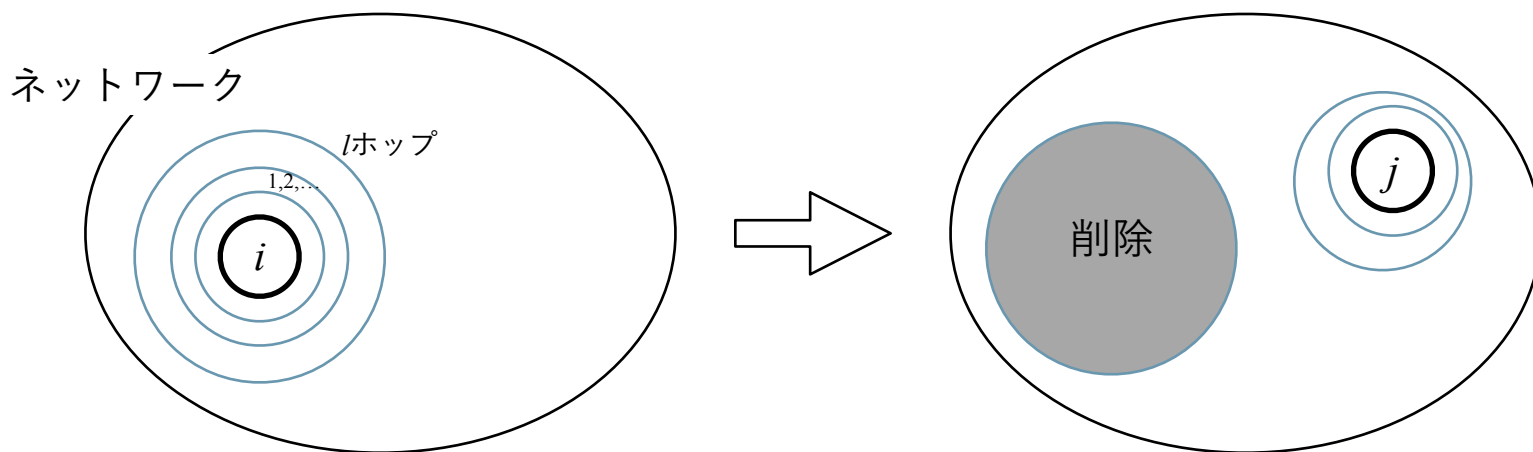
[1]Fuxuan Liao, Yukio Hayashi.

lホップ被覆：離れたシードを選ぶ方法

手順1：全てのノードに対して中心性を計算し、スコアの高いノード i （拡散源となるシード）を一つ選ぶ

手順2：ノード i から l ホップ先まで隣接している全ノードを、ノード i を含めてネットワークから削除する

- ・ ネットワークから全てのノードが削除されるまで以上の処理を繰り返す
- ・ 拡散影響力が高い順番に並んだ i, j, \dots のリストが得られる



より拡散影響力の高いシードを選ぶことが可能

(選ぶことのできるシードの上限は、ネットワークサイズや中心性スコアの手法によって異なる)

拡散源となるノードを選ぶ方法

Katz中心性：あるノードから k ホップ先のノードへと、情報が1ホップ経由するごとに減衰しながら拡散していく現象を捉える

Katz中心性の評価式

$$x_i = \alpha \sum_j A_{ij} x_j + \beta$$

x_i : 中心性スコア
 A_{ij} : 隣接行列

$$\mathbf{x} = \alpha A \mathbf{x} + \beta \mathbf{1}$$

α, β : パラメータ

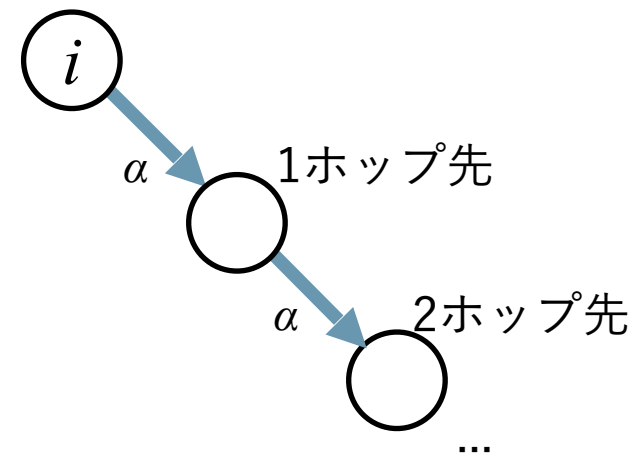
展開すると

$$\begin{aligned} \mathbf{x} &= \beta \frac{1}{1 - \alpha A} \mathbf{1} \\ &= \beta (1 + \alpha A + \alpha^2 A^2 + \alpha^3 A^3 + \alpha^4 A^4 + \dots) \mathbf{1} \\ &= \beta \sum_{k=0}^{\infty} (\alpha A)^k \mathbf{1} \end{aligned}$$

1ホップ経由するごとに α が掛け算される

→ 拡散源から遠いノードほど影響力は小さくなることを再現

→ **拡散影響力の高いノードを選定するには適した手法**



SIRモデルによる拡散シミュレーション

シードがもつ拡散影響力を評価するためのシミュレーション

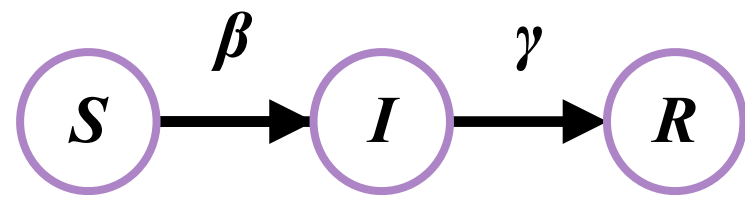
SIR: Susceptible Infected Recoveredモデル

- ・ S (感受性状態), I (感染状態), R (回復状態) の三種類の状態を想定
- ・ ノードを S または I の状態とした初期状態から, 接触感染で感染が伝搬していく時間変化をみる
- ・ 感染率 β で $S \rightarrow I$, 除去率 γ で $I \rightarrow R$ に状態が遷移
- ・ シミュレーションでは $\beta = \lambda \frac{\langle k \rangle}{\langle k^2 \rangle}$, $\lambda = 1, \gamma = 1$ に設定
(SF構造のネットワークでは $\beta = \frac{\langle k \rangle}{\langle k^2 \rangle}$ より大きい値でないと周縁部まで到達しないことが知られている)

常微分方程式で表すSIRモデル

$$\begin{aligned}\frac{dS(t)}{dt} &= -\beta S(t)I(t) \\ \frac{dI(t)}{dt} &= \beta S(t)I(t) - \gamma I(t) \\ \frac{dR(t)}{dt} &= \gamma I(t)\end{aligned}$$

SIRモデルでの状態の遷移



比較する手法・行ったシミュレーション

比較を行う中心性指標

- ・ *l*ホップ被覆を組み合わせた手法

MP-MVC, **HDL** (次数中心性と*l*ホップ被覆), **KCL** (Katz中心性と*l*ホップ被覆)

- ・ 既存手法

HD: 高次数法 (次数中心性), **k-core**: k-core法

LC: 局所中心性, **CI**: 集団的影響力

BC: 媒介中心性, **CC**: 近接中心性

実施したシミュレーション

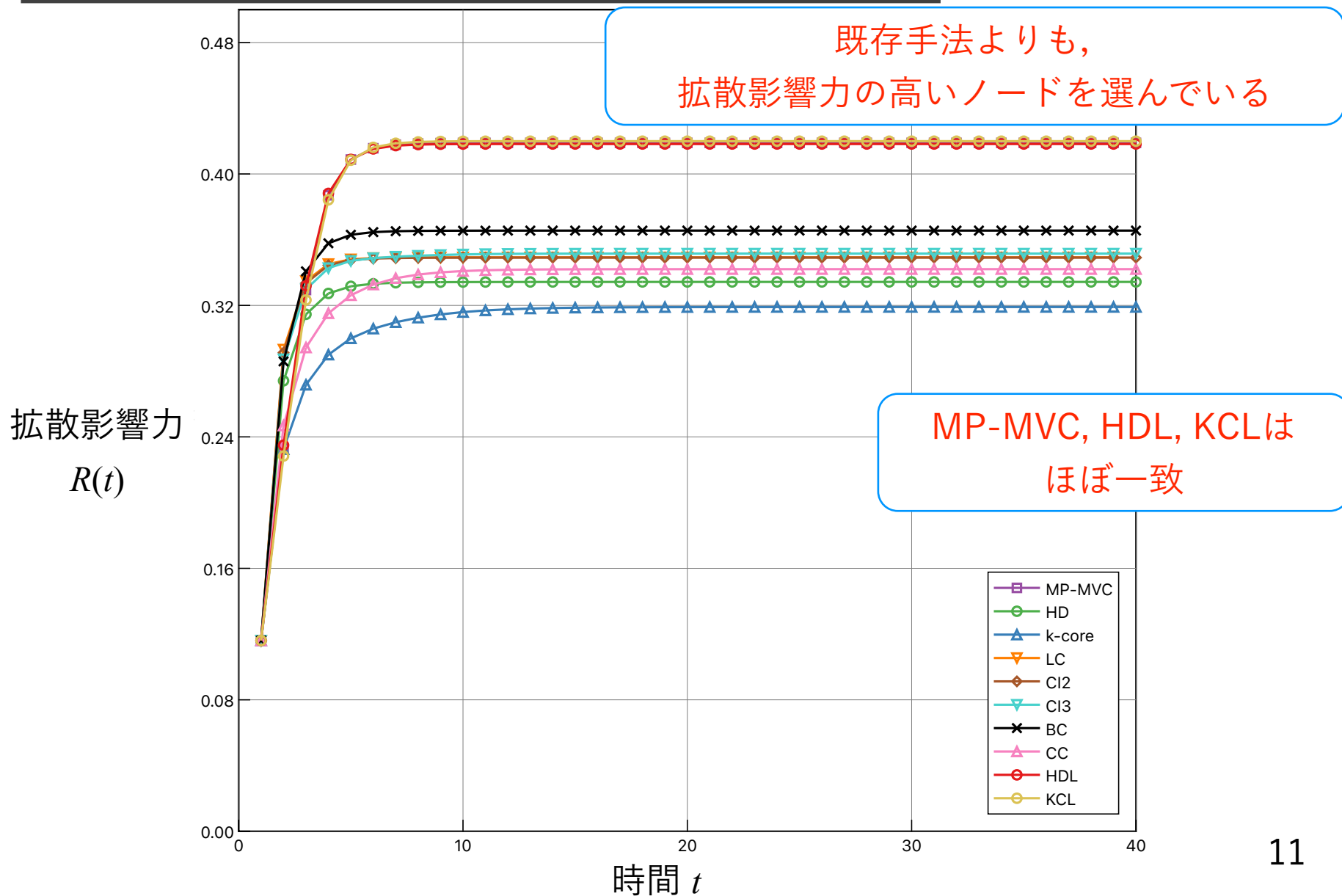
- ・ Scale-Free構造である実データ (LastFM, GRQC, PGP) を使用
- ・ 各手法が選んだシードをグラフ等で比較
 - 拡散影響力の時間変化 (SIRモデルを用いた拡散シミュレーションによる)
 - 距離分布 (シード間の距離)
 - 一致率 (シードの集合としての重なり)

LastFM：各手法でのシード数 N_s

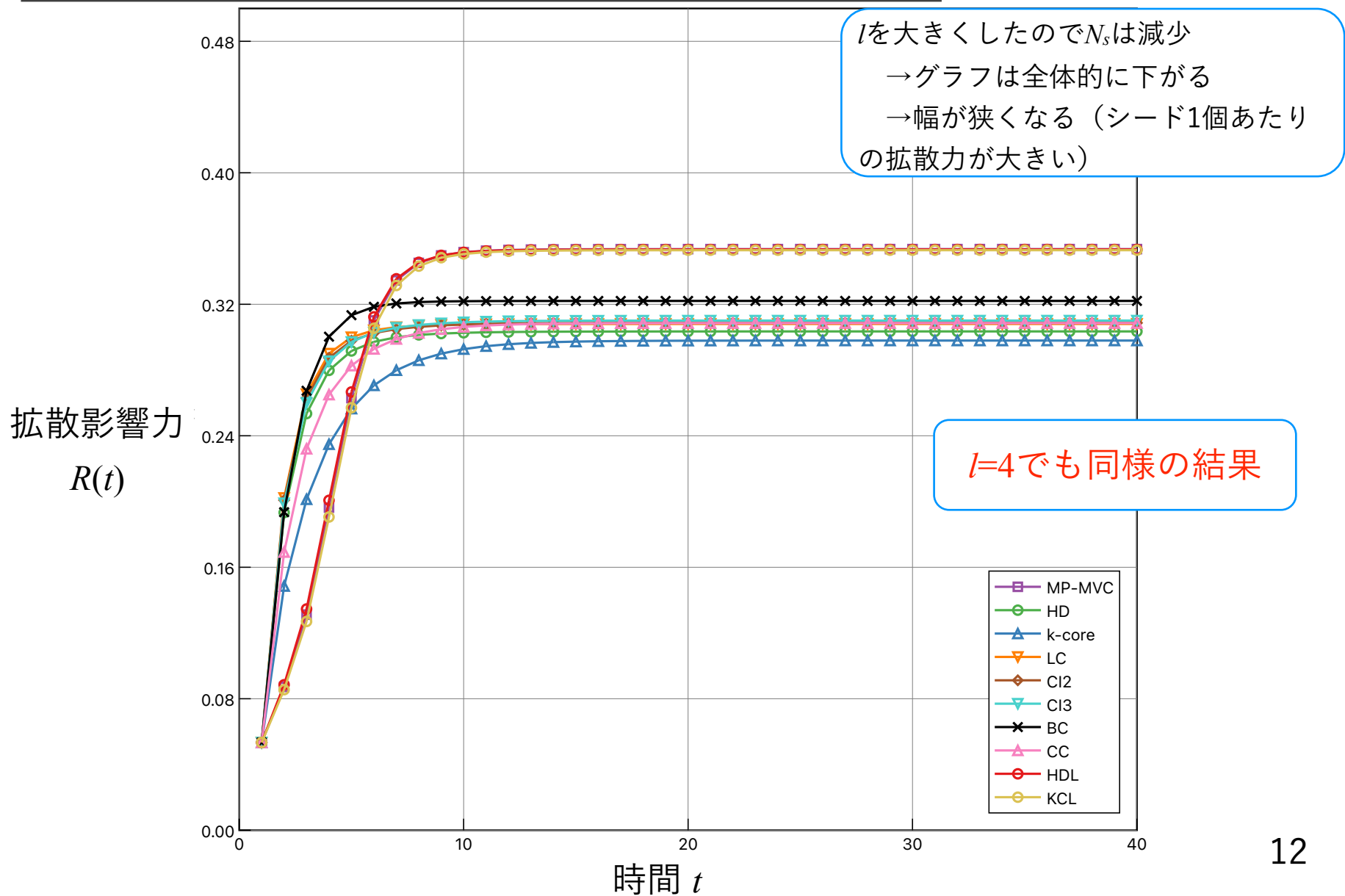
- l ホップ被覆を行う場合には、一番小さい N_s に合わせる必要
(シード上限を超えた値に)
- 結果より次の値に設定
 - $l=2$ のとき, $N_s=885$
 - $l=4$ のとき, $N_s=407$

<u>LastFM</u>	MP-MVC	HDL	KCL
$l=2$	885	2689	2710
$l=4$	407	1656	1661

LastFM：拡散影響力の時間変化($l=2$)

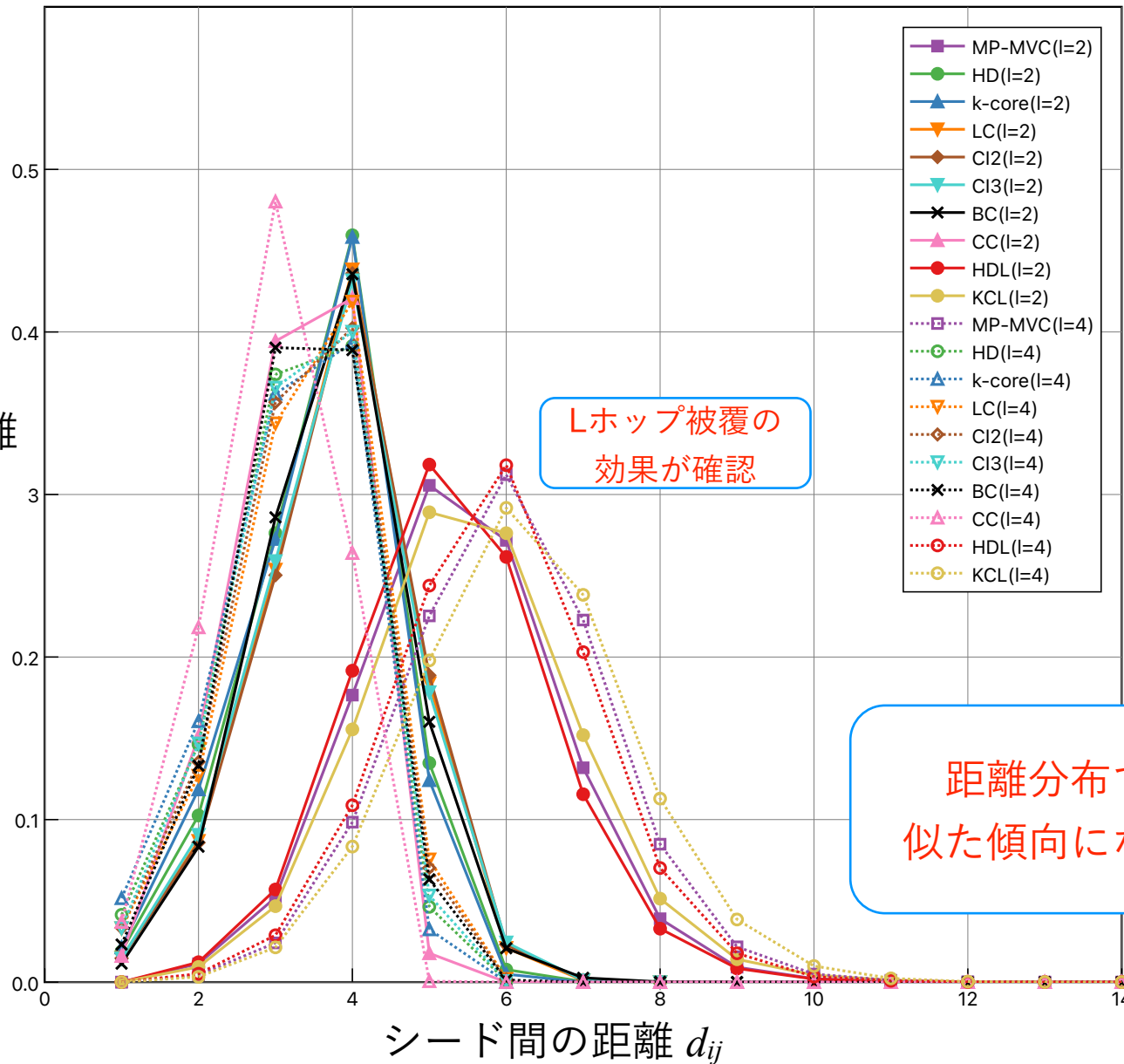


LastFM：拡散影響力の時間変化($l=4$)



LastFM：シード間の距離分布

シード間距離
の頻度
 $P(d_{ij})$



LastFM：各手法が選んだシードの一致率

$l=2, N_s=885$
MP-MVC と HDL：一致数 697
MP-MVC と KCL：一致数 519
HDL と KCL：一致数 433

$l=2$	HDL	KCL
MP-MVC	0.7876	0.5864
HDL	-	0.4893

$l=4, N_s=407$
MP-MVC & HDL：一致数 317
MP-MVC & KCL：一致数 220
HDL & KCL：一致数 166

$l=4$	HDL	KCL
MP-MVC	0.7789	0.5405
HDL	-	0.4079

拡散影響力や、距離分布はほぼ一致
しかし、選んでいるシードは異なった

GRQC, PGP : 各手法でのシード数 N_s

同様に比較を行い, それぞれの値に設定

<u>GRQC</u>	MP-MVC	HDL	KCL
$l=2$	544	1191	1200
$l=4$	306	738	736

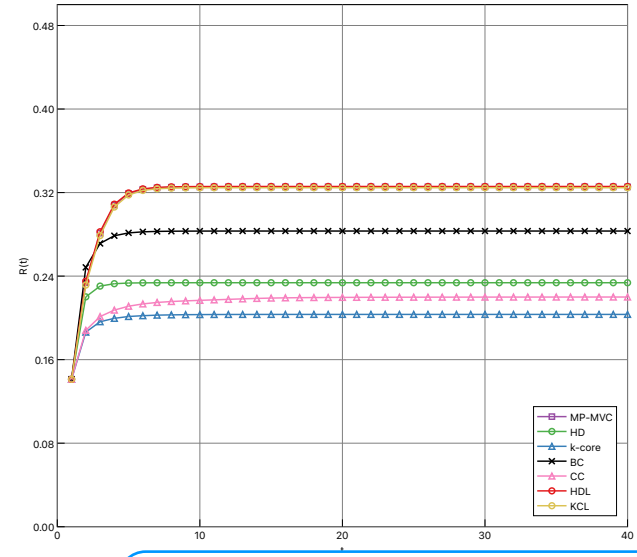
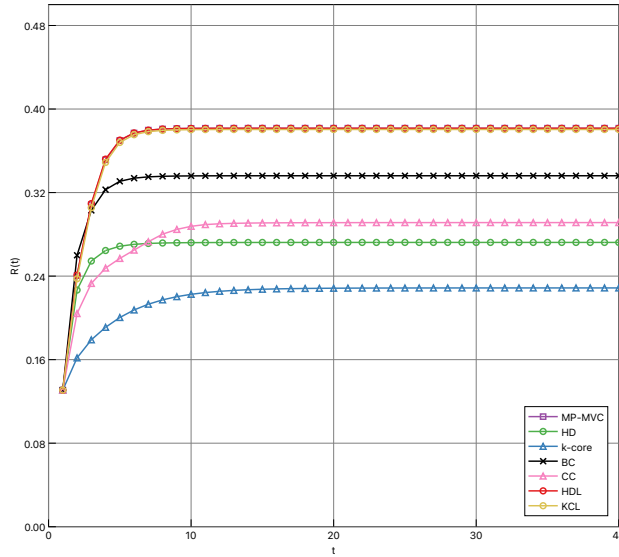
<u>PGP</u>	MP-MVC	HDL	KCL
$l=2$	1508	3593	4025
$l=4$	795	2345	2341

GRQC,PGP : 拡散影響力の時間変化

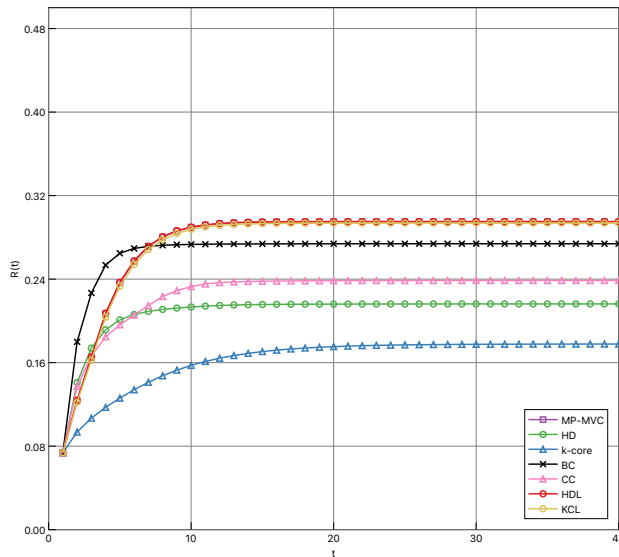
GRQC

PGP

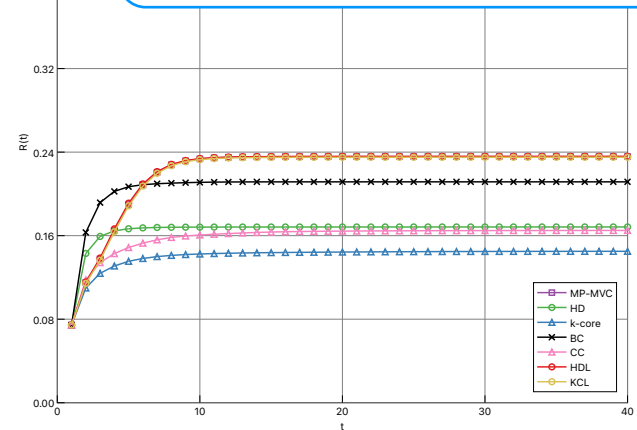
$l=2$



$l=4$



他の二種類の実データでも
同様の結果に



GRQC, PGP : 各手法が選んだシードの一致率

GRQC

$l=2$	HDL	KCL
MP-MVC	0.8456	0.5165
HDL	-	0.4596

PGP

$l=2$	HDL	KCL
MP-MVC	0.8601	0.5637
HDL	-	0.4993

$l=4$	HDL	KCL
MP-MVC	0.8333	0.4935
HDL	-	0.3954

$l=4$	HDL	KCL
MP-MVC	0.8692	0.6101
HDL	-	0.5270

他の二種類の実データでも
シードの一致率は40~85%程度

まとめ

まとめ

- ・ 拡散影響力のより高いシードを選ぶには、シード間の距離が遠いものを選ぶ必要があり、 l ホップ被覆によって実現できる
- ・ 従来手法より優れているMP-MVCと、次数中心性とKatz中心性それぞれに l ホップ被覆を組み合わせて比較した
 - ・ 選んだシードの拡散影響力は同程度だった
 - ・ シード間の距離でも同様な傾向になった
 - ・ しかし、選んでいるシード集合は40～80%程度しか一致していなかった

残された課題

- ・ MP-MVCの原理を理解し、シード集合の一致の要因を考察
- ・ $l=6$ でのシミュレーション
- ・ KCLで、Katz中心性が収束するまで行う（今回は5ホップで打ち切った）

補足：実データの比較

- LastFM：アジアのLastFMユーザーのソーシャルネットワーク
- GRQC：Arxivでの一般相対性理論の共同研究ネットワーク
- PGP：Pretty-Good-Privacyアルゴリズムのユーザーのネットワーク

	$ V $	$ E $	Diameter D	$\frac{\langle k \rangle}{\langle k^2 \rangle}$	N_s at $l=2$	N_s at $l=4$
LastFM	7624	278060	15	0.06	885	407
GRQC	4158	13422	17	0.06	544	306
PGP	10680	24316	24	0.06	1508	795

研究テーマ2

ワイブル分布を用いたネットワークモデルの作成

参考にした論文

[2]Sinan A. Ozbay & Maximilian M. Nguyen

Applied Network Science volume 8, Article number: 20 (2023)

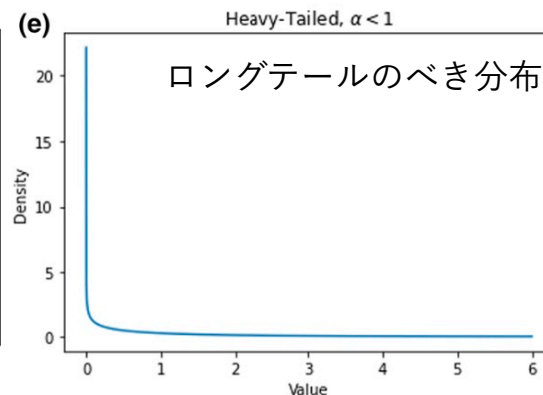
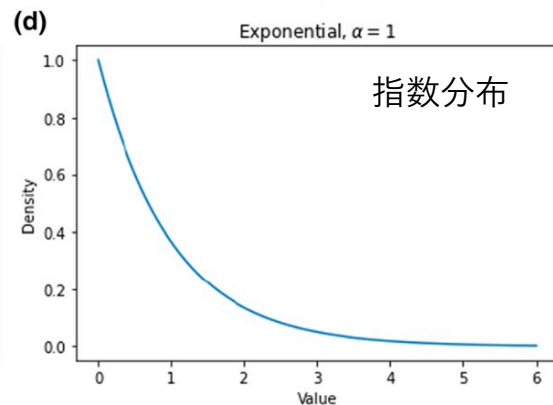
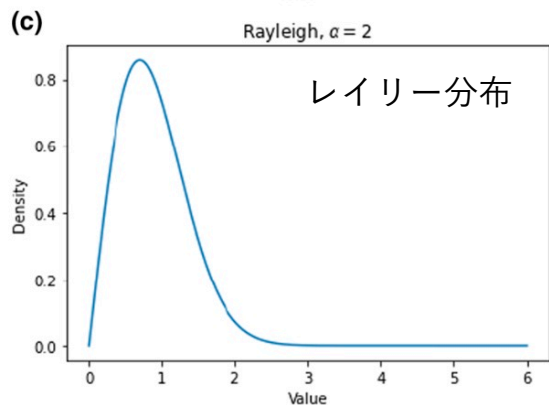
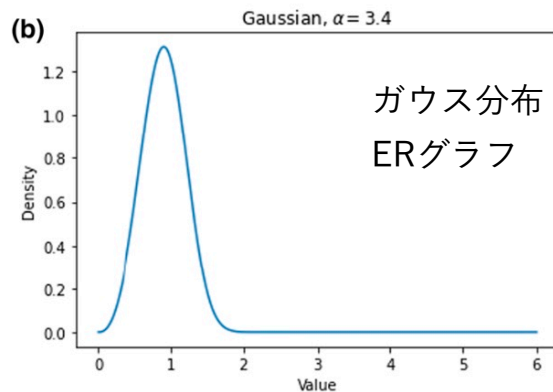
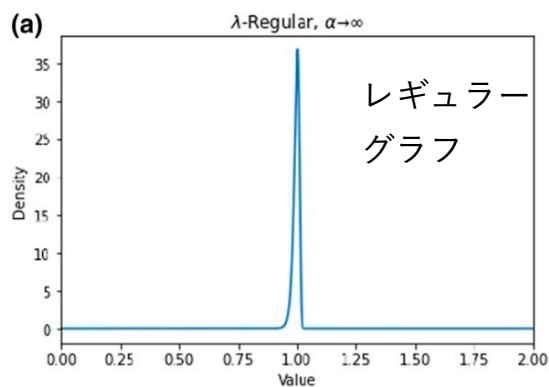
ワイブル分布を用いて次数分布を捉える

ネットワークの次数分布における不均一性を、ワイブル分布（パラメータは α , λ ）を用いて定量的に捉える方法の検討を行う

ワイブル分布の確率密度関数

$$f(x; \lambda, \alpha) = \frac{\alpha}{\lambda} \left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-(x/\lambda)^\alpha}; x \geq 0; \alpha, \lambda > \mathbb{R}^+$$

α : ワイブル係数
 λ : 尺度パラメータ



α の値を変えると、
分布が変化

比較にあたっての条件設定

ワイブル分布と逆優先アルゴリズム (IPA) モデルとを比較

- ・ IPAでのノードがリンクする本数 $m=2$ に合わせる
- ・ ノード数 $N=5,000$, 平均次数 $\langle k \rangle=4$, 最小次数 $k_{min}=2$, 最大次数 $k_{max}=300$

比較するための調整値 C の設定

- ・ 用いたワイブル分布の式 (変数は k, α, λ)

$$f(k) = \frac{\alpha}{\lambda} \left(\frac{k}{\lambda} \right)^{\alpha-1} e^{-(k/\lambda)^\alpha}$$

- ・ C を次式を満たす値に設定 (ただし $f(k)$ を C 倍した後に整数化, それから総和をとった)

$$\langle k \rangle N = 2M$$

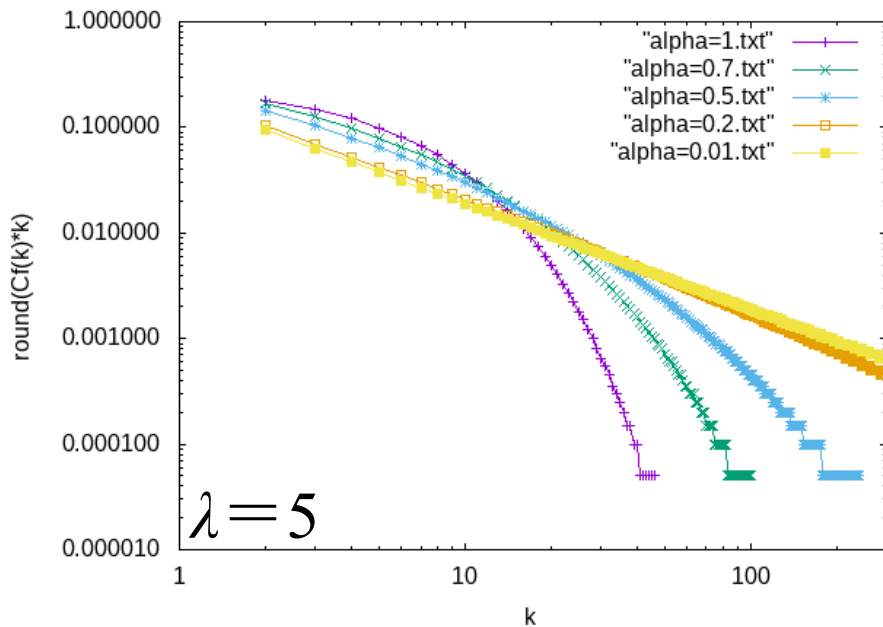
$$2M = \sum_{k=k_{min}}^{k_{max}} C f(k)$$

α を変えた際の変化 ($\lambda=5$)

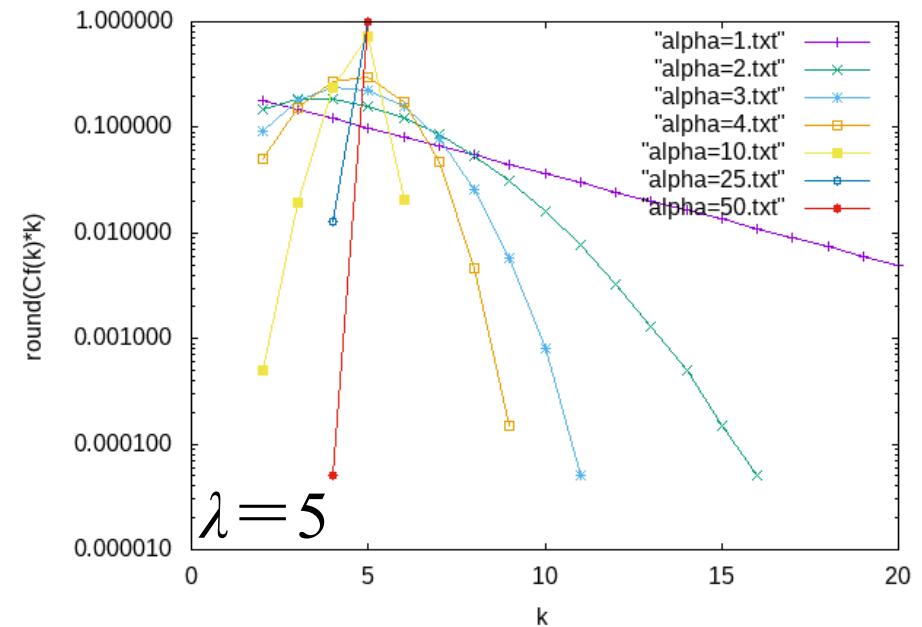
α を変えるとどうなるか (縦軸は $Cf(k)$ を整数化した値を $2M$ で割った値)

- α を大きくしていくと分布は狭くなっていく
- 逆に小さくしていくと右側に膨らんでいき、直線 (平坦) に近づく

Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=5)



Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=5)

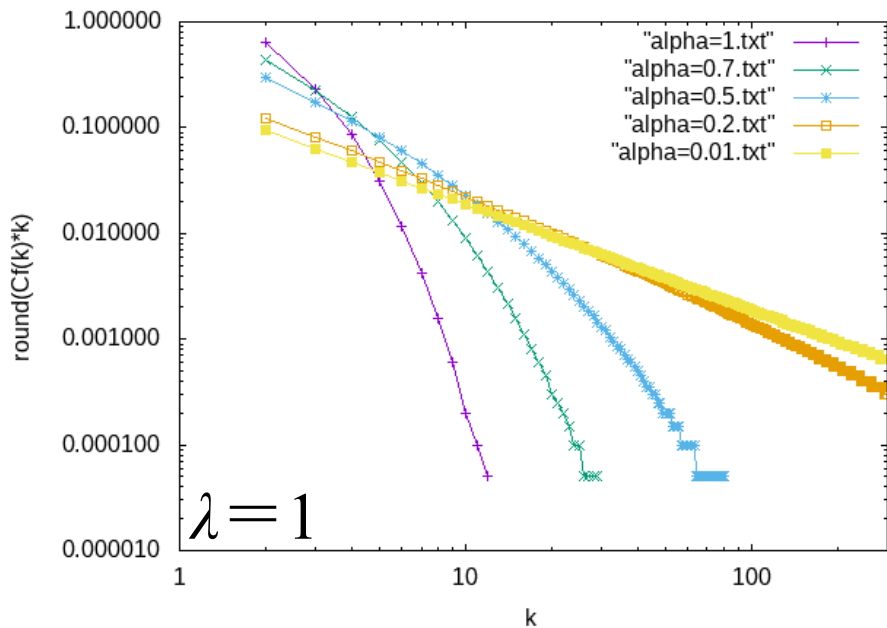


α を変えた際の変化 ($\lambda=1$)

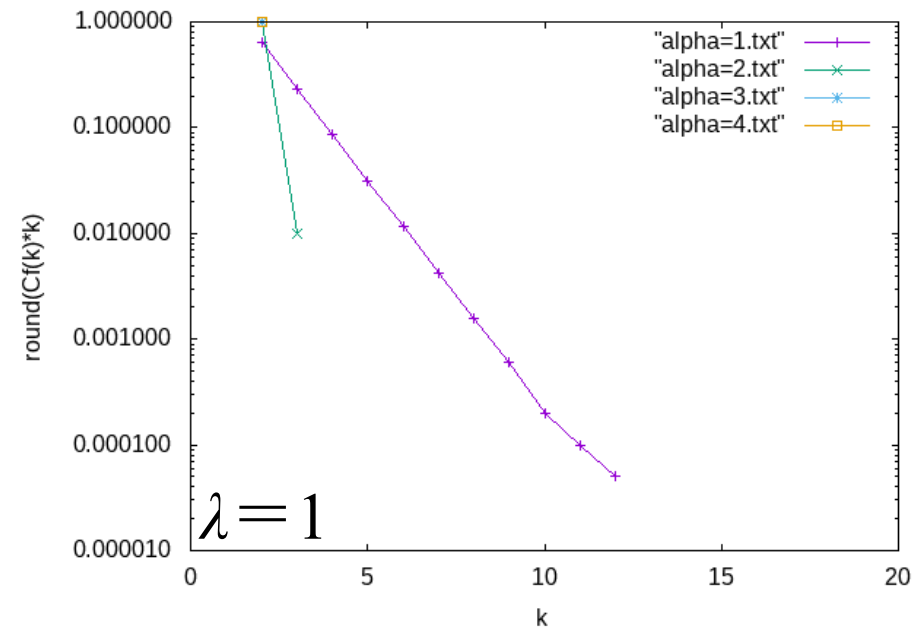
α を変えるとどうなるか (縦軸は $Cf(k)$ を整数化した値を $2M$ で割った値)

- α を大きくしていくと分布は狭くなっていく
- 逆に小さくしていくと右側に膨らんでいき、直線 (平坦) に近づく

Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=1)



Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=1)



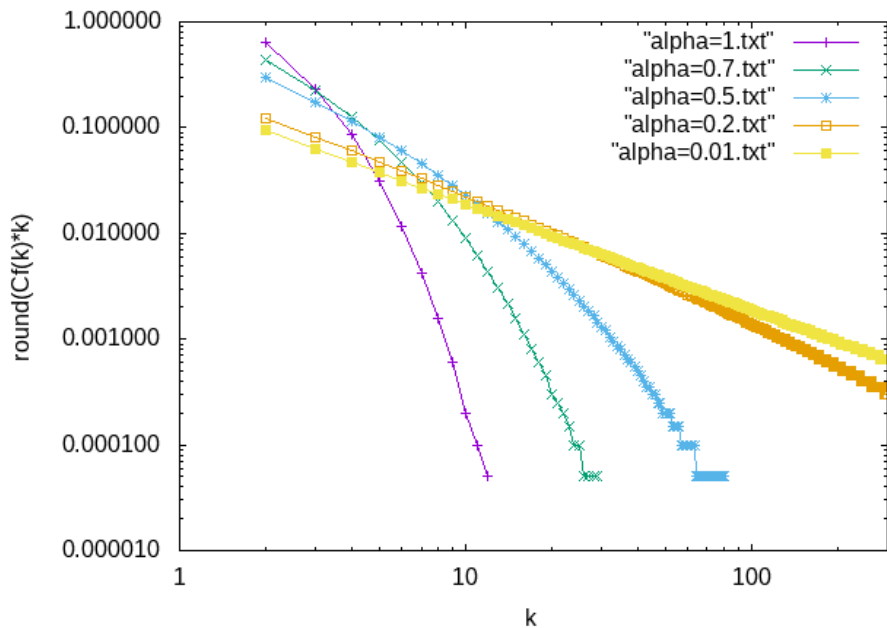
λ を変えた際の変化 ($\alpha \leq 1$)

λ を変えるとどうなるか (縦軸は $Cf(k)$ を整数化した値を $2M$ で割った値)

- λ を大きくすると、左端がやや下がりながら右側に膨らんでいく

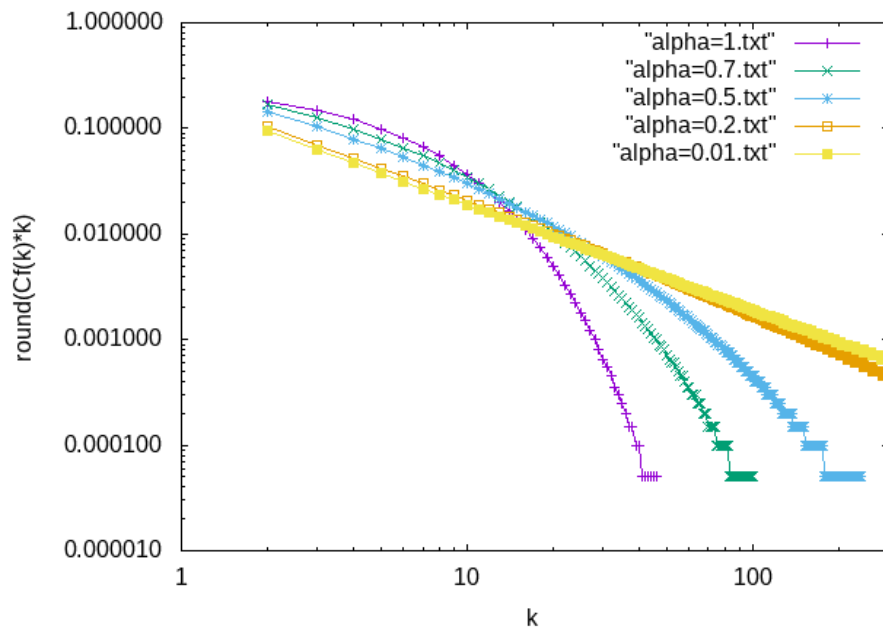
$\lambda=1$

Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=1)



$\lambda=5$

Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=5)



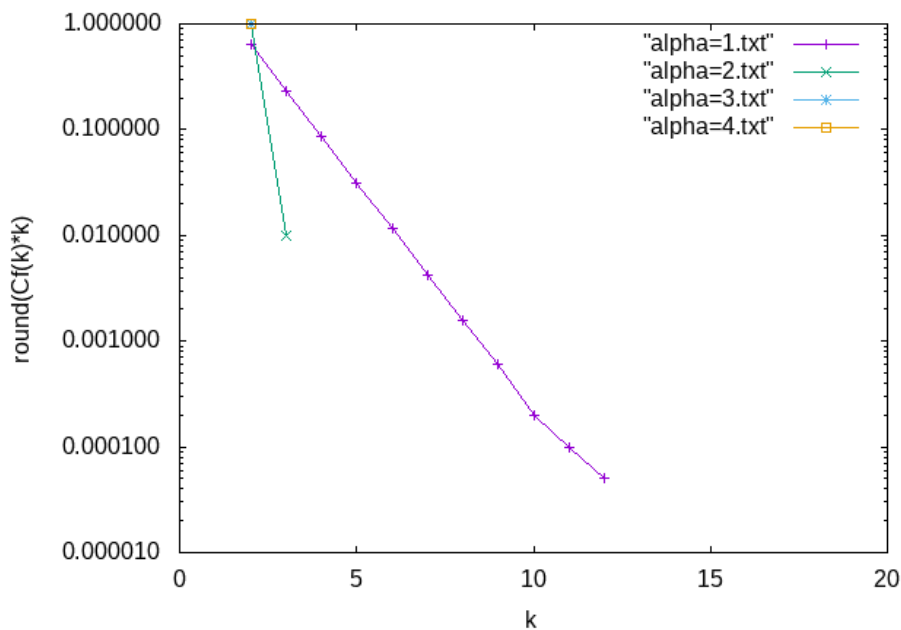
λ を変えた際の変化 ($\alpha \geq 1$) ①

λ を変えるとどうなるか (縦軸は $Cf(k)$ を整数化した値を $2M$ で割った値)

- λ を大きくすると、左端がやや下がりながら右側に膨らんでいく

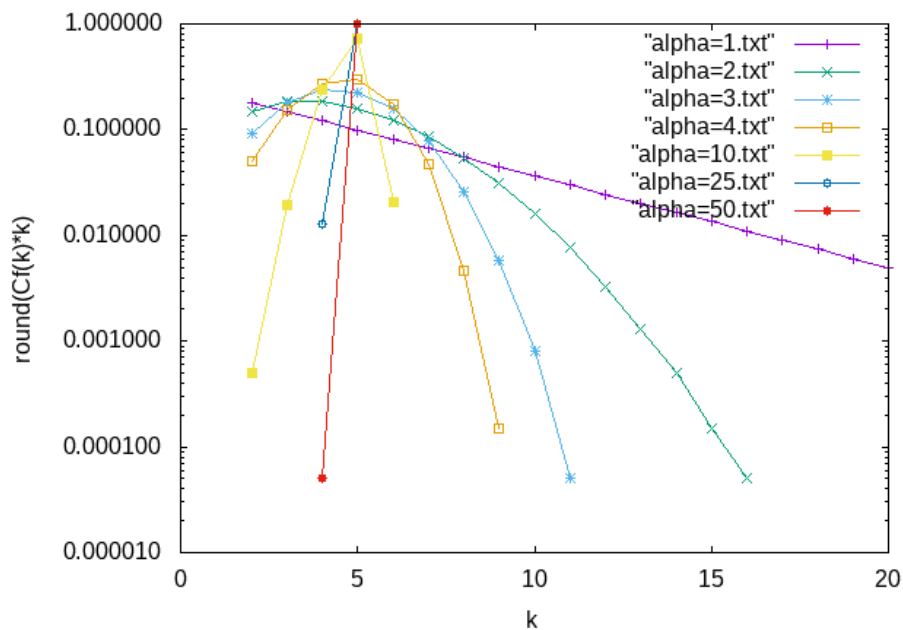
$\lambda=1$

Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=1)



$\lambda=5$

Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=5)



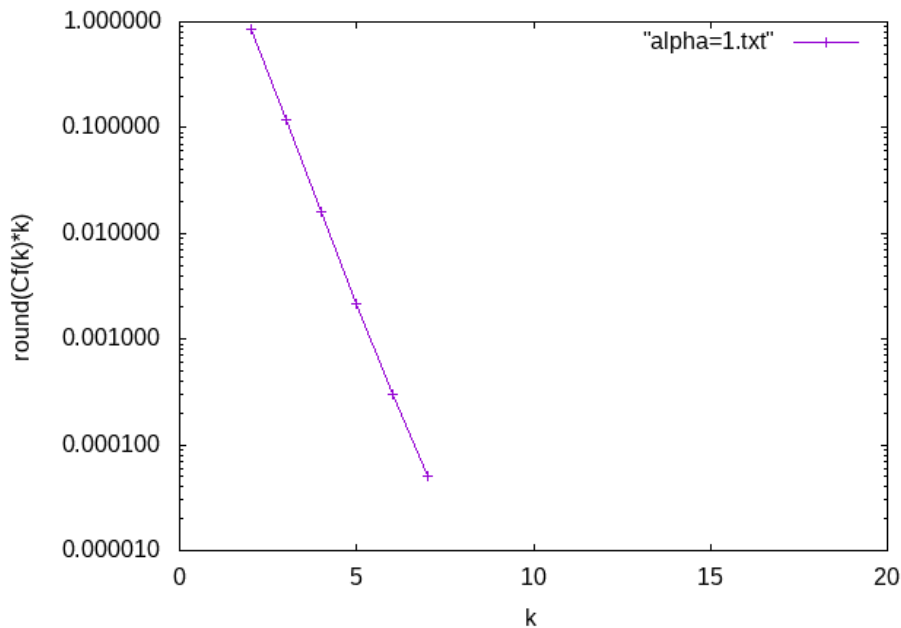
λ を変えた際の変化 ($\alpha \geq 1$) ②

λ を変えるとどうなるか (縦軸は $Cf(k)$ を整数化した値を $2M$ で割った値)

- λ を大きくすると、左端はほぼ変わらずに右側に膨らんでいく

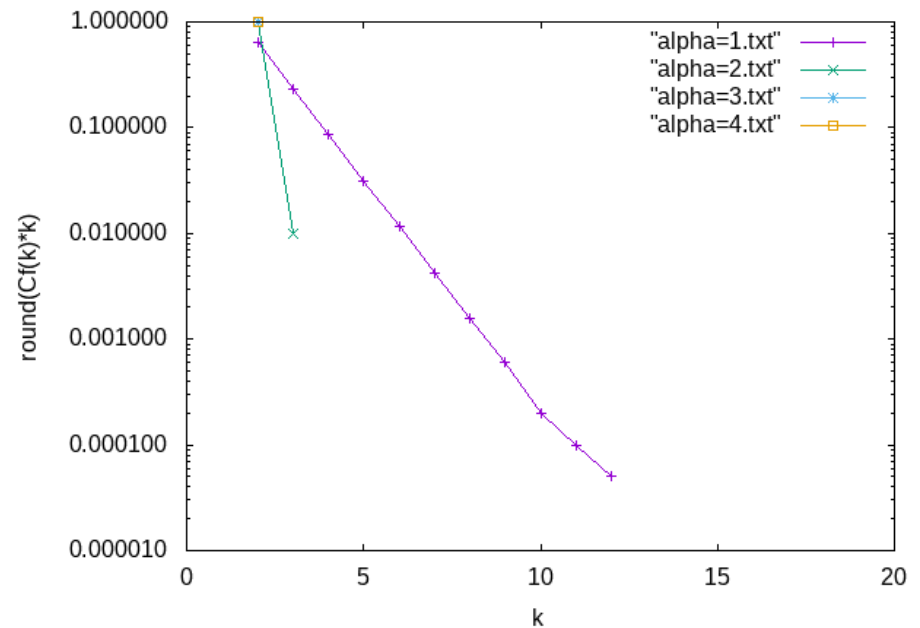
$\lambda=0.5$

Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=0.5)



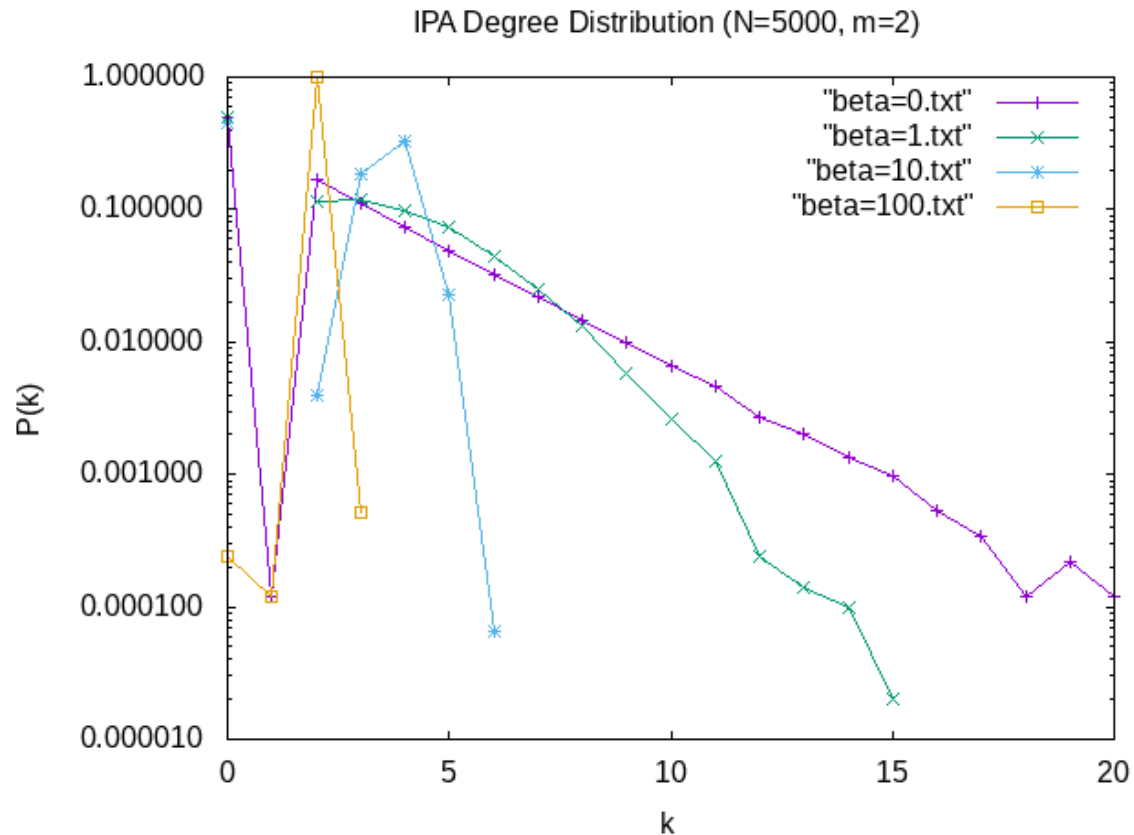
$\lambda=1$

Weibull Distribution (N=5000, <k>=4, kmin=2, kmax=300, lambda=1)



IPAモデルの次数分布

- $N=5000$, $m=2$ に固定し, $\beta = 0, 1, 10, 100$ として, 5回分の平均をプロット
- β が大きくなるにつれて分布の幅が狭くなっていく様子が見られる

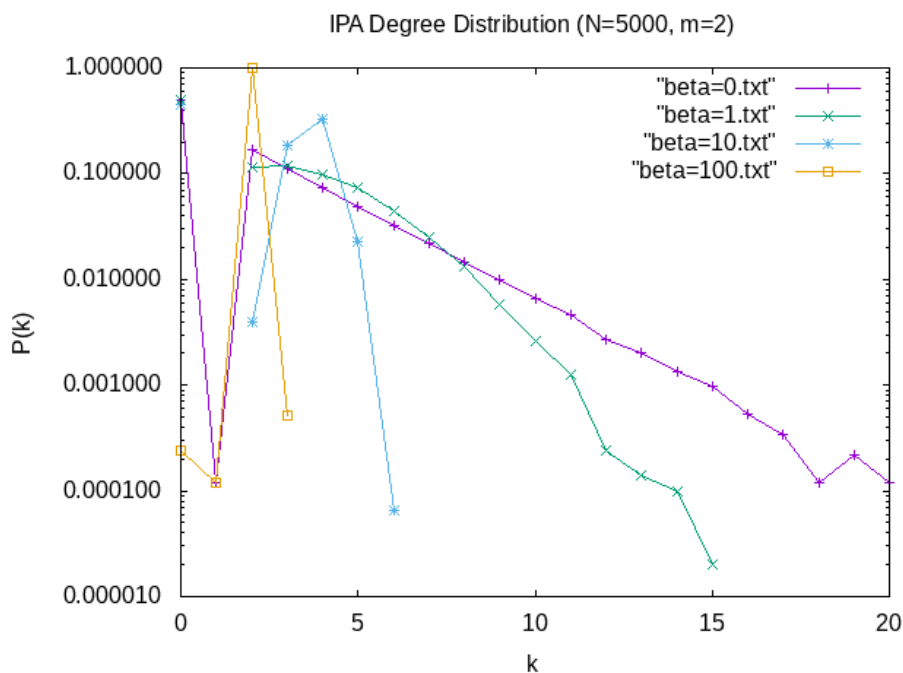


ワイブル分布とIPAの次数分布を比較

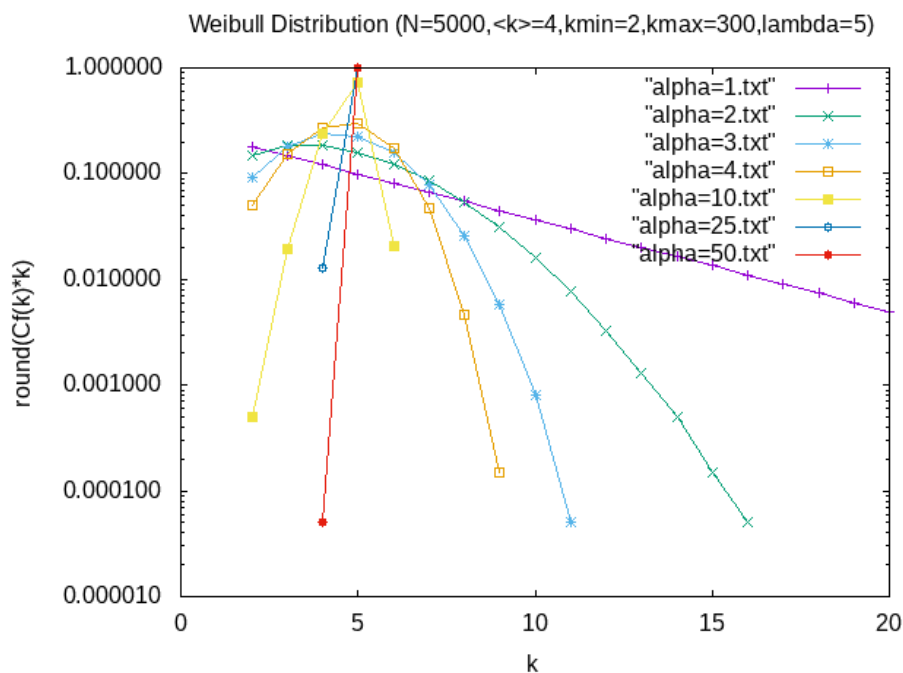
近い傾向にあるグラフ（目視による）

- ・ IPAとワイブル分布でそれぞれ、 $\beta=0$ と $\alpha=1$ 、 $\beta=1$ と $\alpha=2$ 。
- ・ IPAモデルの方が分布の幅が狭い

IPAモデルの次数分布 ($m=2$)



ワイブル分布 ($\lambda=5$)



(縦軸は $Cf(k)$ を整数化した値を $2M$ で割った値)

まとめ

まとめ

- ・ワイブル分布のパラメータ α と λ をそれぞれ変えた際のグラフの変化を、目視で確認し、IPAモデルの次数分布との比較を行った
- ・ α と λ の値の変化によって、分布がどのように変化するかを確認できたが、IPAモデルとどの程度類似しているかはさらに検討が必要

残された課題

- ・比較できるように今回作成したワイブル分布を次数列に置き換え、コンフィグモデルに与えてネットワークモデルを作成する
- ・作成したネットワークモデルの次数分布とIPAモデルの次数分布を比較する
- ・グラフの類似度を定量的に評価する

難航した要因

- ・コンフィグモデルを作成することはできたが、与える次数列によってエラーが発生した（次数分布に対して、与える次数列への変換のプログラムがうまく作成できなかった）
- ・ワイブル分布での α や λ の変化によってCの値が大きく変化して予測できず、組み合わせによっては時間がかかった（ほぼ手動で調整していたため、補足を参照）

補足：パラメータ C の値

$N=5000, \langle k \rangle=4, k_{min}=2, k_{max}=300$ の場合

• $\lambda=5$ に固定

$\alpha=1: C=27047.1150, \alpha=2: C=21824.3460, \alpha=3: C=20487.9600, \alpha=4: C=20127.6100,$

$\alpha=10: C=19636.1890, \alpha=25: C=10735.3110, \alpha=50: C=5436.4270, \alpha=100: C=2718.2140,$

$\alpha=0.01: C=1030019.0990, \alpha=0.05: C=207118.5020, \alpha=0.1: C=105490.3340,$

$\alpha=0.2: C=57013.2970, \alpha=0.5: C=34761.0620, \alpha=0.7: C=30866.4449$

• $\lambda=1$ に固定

$\alpha=1: C=93416.1410, \alpha=2: C=270254.0180, \alpha=3: C=4968139.1060, \alpha=4: C=5553900000,$

$\alpha=0.01: C=1030391.4770, \alpha=0.2: C=68273.8490, \alpha=0.5: C=68914.1699, \alpha=0.7: C=77040.6749,$

• $\lambda=0.5$ に固定

$\alpha=1: C=472096$

(小数点以下4桁までを算出するようにしたが、 $f(k)$ に C を掛け算した後に整数化してから総和をとるので、1桁でもよかったのでは)