READING:

The Power of Random Neighbors in Social Networks

Silvio Lattanzi
Google, Inc.
New York, NY 10011
silviol@google.com

Yaron Singer
Harvard University
Cambridge, MA 02138
yaron@seas.harvard.edu

Speaker: Kazuhiro Inaba

Paper Introduction from WSDM 2015

背景: “Friendship Paradox” [Feld 91]

「Social Network において、（平均的に見ると）
本人よりも、その**友人**の方が、友達が多い。」

$$\mu = \Sigma_{v \in V}\, d(v) \,/\, |V|$$
$$= 2|E| \,/\, |V|$$

$$\leqq$$

$$\Sigma_{(u,v) \in E}\, (d(u)+d(v)) \,/\, 2|E|$$
$$= \Sigma_{v \in V}\, d(v)^2 \,/\, 2|E|$$
$$= \Sigma_{v \in V}\, d(v)^2 \,/\, |V| \,/\, \mu$$
$$= \mu + \sigma/\mu$$

この論文の内容

- **Friendship Paradox in a “Strong” sense.**
  - ある自然なモデルにおいて、
    |友人の平均次数| / |平均次数| ∈ ω(1)
    つまりグラフを大きくすると友人との格差が
    より開く現象が観測できることを示す。

- 実際のグラフデータで実験 (??)
- グラフアルゴリズムへの影響・示唆 (??)

Section 3: Misbehaved Power Laws

- **Prop 3.1 次数の冪分布 だけでは asymptotic friendship paradox が成り立たないことがある**

- 証明: 冪指数 β=2 の例を構成。
  - 次数の小さい順に頂点を列挙し、辺を貼っていく
  - 次数 (√N) / 400 までは貪欲に、列挙順が自分に近い頂点と結ぶ
  - 残りは適当に結ぶ

【論文を通して用いるメタ変数の定義】
N: 頂点数　M: 辺数　β: べき指数　C: 次数1の頂点の数
(次数 d の頂点の数は C / d^{β},　最大次数は $C^{1/\beta}$)

Section 4: Single Samples ($\beta < 2$)

- $G(\beta, p)$ を、次数が冪指数 β の冪分布のグラフの各辺を小さい確率 p でランダムにつなぎ替えたグラフモデルとする

- **Prop 4.1**
  $1<\beta<2$ なら、$G(\beta,p)$ のグラフから頂点 v をランダムにとり、その隣接頂点 u をランダムにとったとき、任意のεに関し、ある定数以上の確率で $d(u)/d(v) \in \Omega(N^{(\beta-1)/\beta-\varepsilon})$.

Section 4: Single Samples
(証明の前半: 乱択した時の次数は高くない)

- 頂点数 N は C と β に対してこのくらい

$$N = \sum_{i=1}^{\Delta} \left\lfloor \frac{C}{i^{\beta}} \right\rfloor = \sum_{i=1}^{\lceil C^{1/\beta} \rceil} \left\lfloor \frac{C}{i^{\beta}} \right\rfloor \in \Theta(C)$$

- 辺数 M は C と β に関して

$$M = \sum_{i=1}^{\lceil C^{1/\beta} \rceil} i \left\lfloor \frac{C}{i^{\beta}} \right\rfloor = \Theta\left(\sum_{i=1}^{\lceil C^{1/\beta} \rceil} \left\lfloor \frac{C}{i^{\beta-1}} \right\rfloor \right) \in \Theta\left(C^{\frac{2}{\beta}} \right)$$

∵ Σ^x i^{1-β} ≃ x^{2-β}

- よって平均次数が $\Theta\left(C^{\frac{2}{\beta}-1}\right)$ で、Markovの不等式より
  確率 1/k で平均次数の k 倍以下のノードに当たる。

Section 4: Single Samples
(証明の後半: 乱択した時の隣人の次数は高い)

- 次数 K 以上の頂点の個数は

$$A_{d>K} = \sum_{i=\lceil K \rceil}^{\lceil C^{1/\beta} \rceil} i \left\lfloor \frac{C}{i^\beta} \right\rfloor = \Theta\left(C\left(C^{\frac{2}{\beta}-1} - K^{2-\beta}\right)\right)$$

- $K = C^{1/\beta - \epsilon}$ とするとこれは $\Theta(M)$
- 確率 p で rewire された辺が繋がっている次数 K 以上の頂点の個数も $pA_{d>K} \in \Theta(M)$
- よって一定の確率で
隣人の次数 $K = C^{1/\beta - \epsilon}$ / 本人の次数 $\Theta\left(C^{\frac{2}{\beta}-1}\right)$
という比になる。つまり

$$\Omega\left(C^{\frac{\beta-1}{\beta}-\epsilon}\right) = \Omega\left(N^{\frac{\beta-1}{\beta}-\epsilon}\right)$$

Section 4. Single Samples (β>2)

- **Prop 4.2.**
  2 < β < 3 では、G(β,p) のグラフから頂点 v をランダムにとり、その隣接頂点 u をランダムにとるとき、ほとんど常に d(u)/d(v) ∈ Ω(1).
- **違いは:** $\int i^{1\text{-}\beta} = i^{2\text{-}\beta} / (2\text{-}\beta)$

β<2 のとき $$A_{d>K} = \sum_{i=\lceil K\rceil}^{\lceil C^{1/\beta}\rceil} i \left\lfloor \frac{C}{i^{\beta}} \right\rfloor = \Theta\left(C\left(C^{\frac{2}{\beta}-1} - K^{2-\beta}\right)\right)$$

β>2 のとき $$A_{d>K} = \sum_{i=\lceil K\rceil}^{\lceil C^{1/\beta}\rceil} i \left\lfloor \frac{C}{i^{\beta}} \right\rfloor = \Theta\left(C\left(K^{2-\beta} - C^{\frac{2}{\beta}-1}\right)\right)$$

よって K∈ω(1) なら $A_{d>K}$ ∈ o(N) なので、
ほとんどの点は次数K以上の辺に接していない

Section 5. Poly-Logarithmic Samples

- **Prop 5.2**
  $2 < \beta < 3$ では、G(β,p) のグラフからサイズ log N の頂点集合 S をランダムにとり、Sの各点からランダムに隣接点を取った集合を T_S とすると、高い確率で $d(T_S)/d(S) \in \omega(1)$.

2 < β の G(β,p) モデルでは、

・ 1人の人間に関する “Asymptotic” Friendship Paradox は発生しない
・Polylog人の集団を集めてくると発生する

Section 5. Poly-Logarithmic Samples
(証明の大雑把な直感)

次数 K = log N 以上の点に接している辺の数は以下の通りで

$$A_{d>K} = \sum_{i=\lceil K \rceil}^{\lceil C^{1/\beta} \rceil} i \left\lfloor \frac{C}{i^{\beta}} \right\rfloor = \Theta\left(C\left(K^{2-\beta} - C^{\frac{2}{\beta}-1}\right)\right)$$

(再掲)

辺の総数は以下の通りなので

$$M = \sum_{i=1}^{\lceil C^{1/\beta} \rceil} i \left\lfloor \frac{C}{i^{\beta}} \right\rfloor = O\left(\sum_{i=1}^{\lceil C^{1/\beta} \rceil} \left\lfloor \frac{C}{i^{\beta-1}} \right\rfloor\right) = \Theta(C)$$

K 人集めてくると $K A_{d>K} / M =$ **$K^{3-\beta}$** 人くらいは
次数>K の隣人が入ると期待される。故に平均次数≧**$K^{3-\beta}$**.

Section 6. 実験

- 実際のネットワークで d(u)/d(v) 比はどうなっているか。

データセット:

$Network$	$\#\ of\ Nodes$	$\#\ of\ Edges$	β	C
Orkut	3,072,441	117,185,083	0.7470	223,989
LiveJournal	3,997,962	34,681,189	1.0322	520,041
Wikipedia	2,394,385	5,021,410	1.9548	80,033
YouTube	1,134,890	2,987,624	1.4212	160,927
DBLP	317,080	1,049,866	1.2048	64,983
SlashDot	82,168	948,464	1.2146	13,805
Enron	36,692	367,662	1.1636	11,322

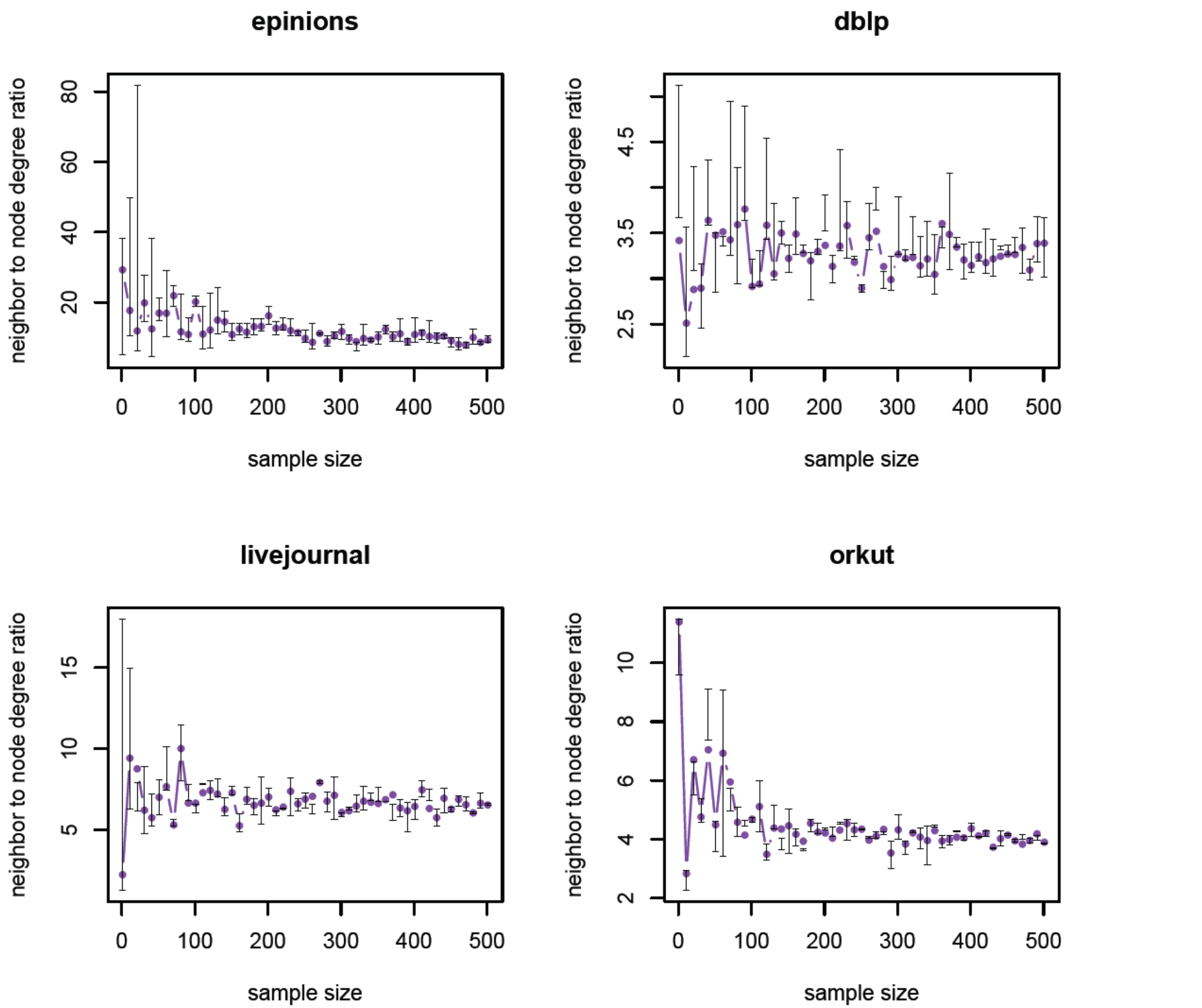
Section 6. 実験（平均次数の比）

横軸 =

とった
サンプルの
サイズ

縦軸 =

隣接集合の
平均次数
/ サンプルの
平均次数



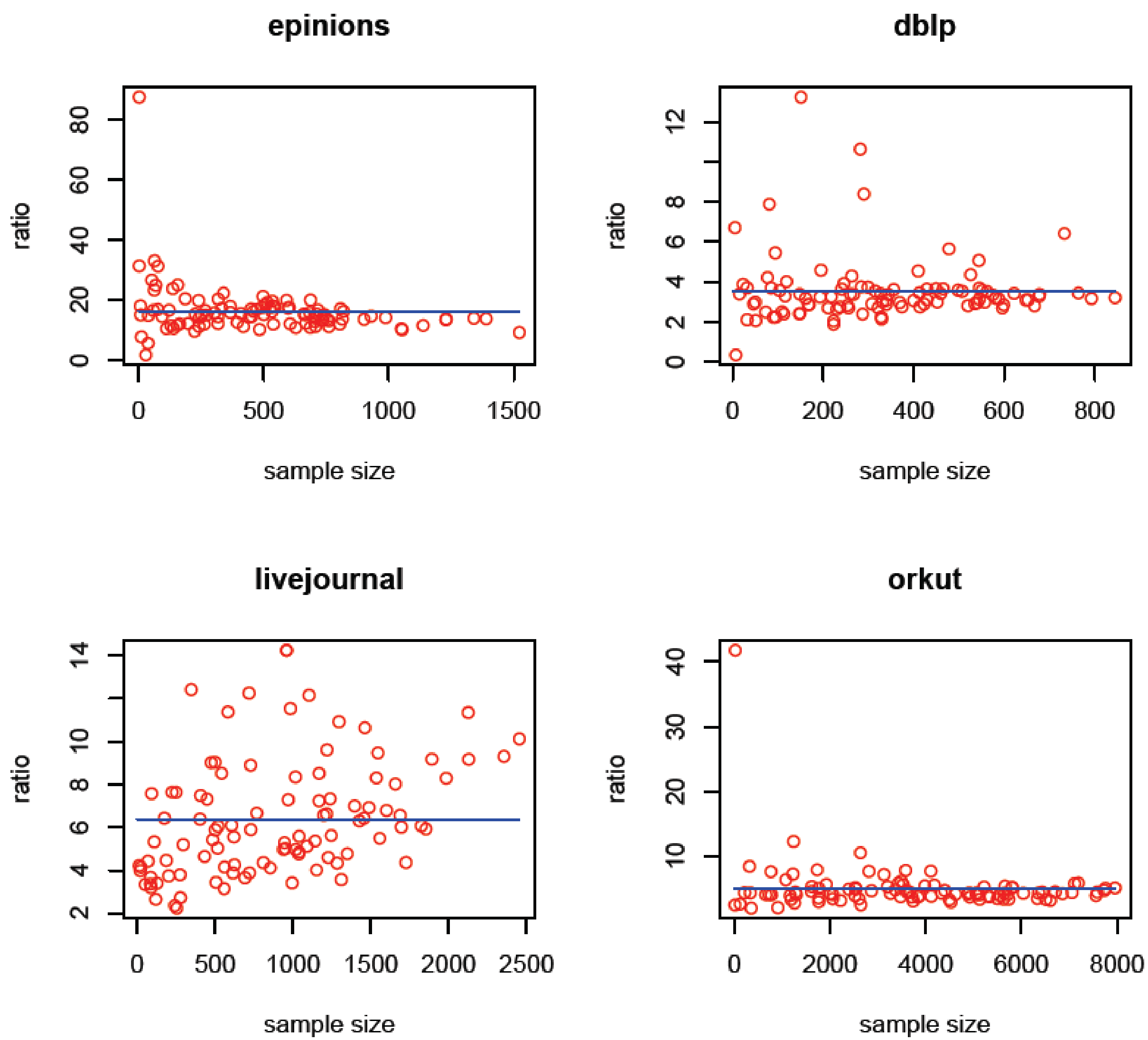
Section 6. 実験（平均次数の比）

横軸 =

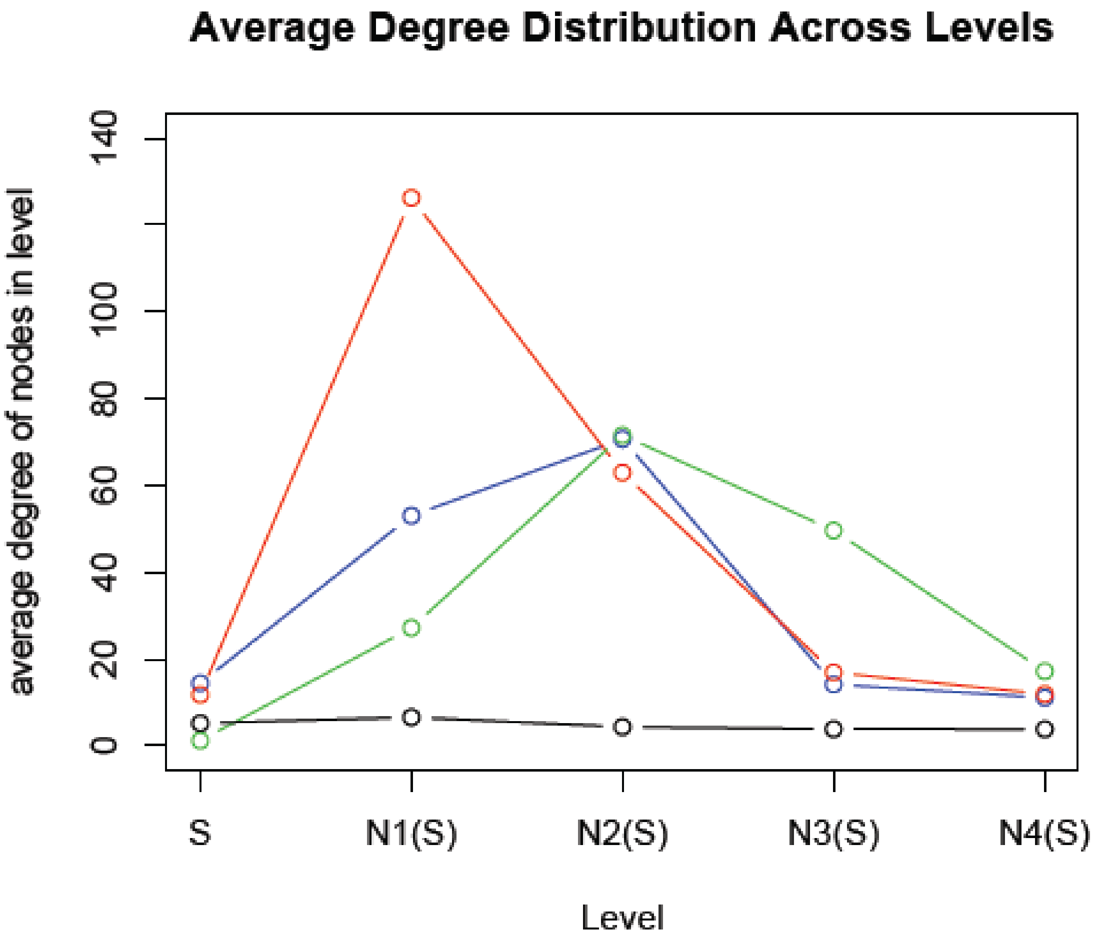
とった
サンプルの
サイズ

縦軸 =

隣接集合の
平均次数
/ 隣接集合と
同じサイズの
乱サンプルの
平均次数



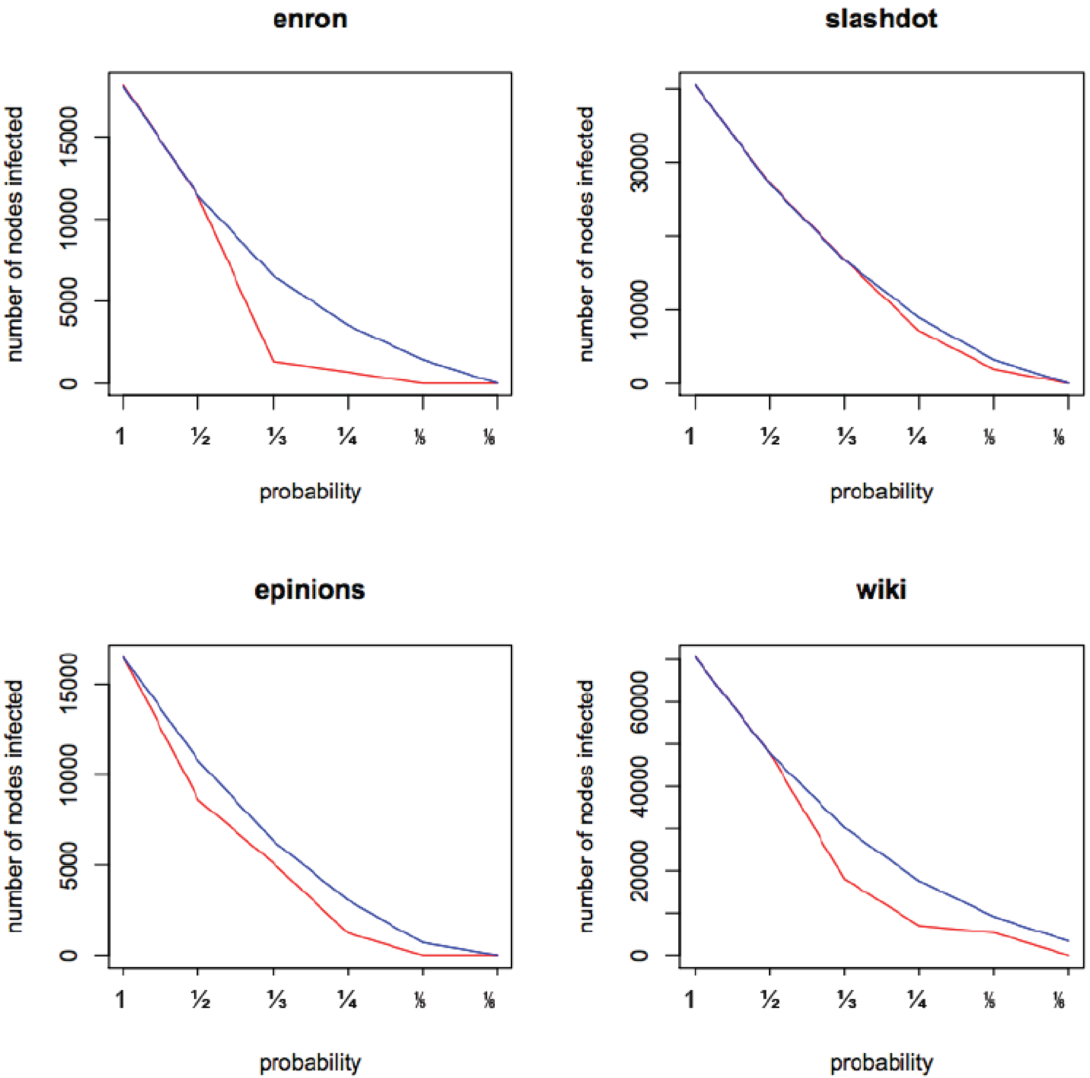
Section 6. 実験（友人の友人）



Section 6. 実験（Influence Maximization）
(多分 Independent Cascade)

赤線 =
- 10 random node が seed

青線 =
- 10 random neighbor が seed



まとめ

- ランダムにサンプルした頂点よりもその隣接ノードの平均次数の方が高いことが多い (Friendship Paradox)
- 次数が冪分布なモデルでは、この次数比率はグラフのサイズ N に依存して大きくなる
  - $\beta < 2$ と $2 < \beta$ での違い
- 実験が少ない？
  - $2 < \beta < 3$ の実データ？
  - Asymptotic Behavior の実験的検証？