補足ノート 時田恵一郎（名大情報）(Sep. 11, 2014)

M. P. H. Stumpf, C. Wiuf and R. M. May
Subnets of scale-free networks are not scale-free: Sampling properties of networks
スケールフリーネットワークの部分ネットはスケールフリーではない：ネットワークのサンプリング特性
PNAS, **102** (2005) 4221-4224.

◎Robert May はランダム行列理論による生態系の複雑さと安定性の関係 [1]，ロジスティックカオスの発見 [2]，感染症 [3] やウィルス [4] の進化のダイナミクスなどで知られる，数理生物学，理論生物学の創始者の一人．前王立協会（Royal Soceity）会長．

概要

一般的なネットワークに対する「二項（binomial）二段サンプリング（subsampling: 孤児（orphan）ノード除去）」について，無限サイズのネットワーク（母集団）と部分ネット（サンプル）の確率母関数の間に成り立つ定理を導出し，スケールフリーネットワークの場合には母集団とサンプルの確率母関数が一致しないこと，および負の二項分布に関してはそれらが一致することを示した．

ネットワーク $\mathcal{N}$（サイズ，またはノード数 N，次数分布 $P(k)$）を考える．サイズはあとで無限大（$N \to \infty$）を考える（有限サイズの母集団からのサンプリングはややこしいので）．

次数分布のみでネットワークを特徴付けられるわけではないことには注意．全く異なるネットワーク（ループ有り＆木, etc）が同じ次数分布をもつことがある．しかし，次数分布はネットワークを特徴付ける量の一つとして最もよく調べられており（他にはクラスタリング係数やネットワークの直径など），一般に次数分布がべき分布になっていることがスケールフリー性の判定に用いられるので，ここでは次数分布に対するサンプリングの効果に注目する．

二項サンプリング (binomial sampling)

各ノードを確率 $p(0 \le p \le 1)$ でサンプルする．サンプルされたノードからなる部分ネットを $\mathcal{S}$ とする．部分ネット $\mathcal{S}$ のサイズを M とすると，M は明らかに二項分布

$$\mathrm{Bi}(M) \equiv {}_N C_M p^M (1-p)^{N-M} \tag{1}$$

に従い，その期待値と分散（付録 A）は

$$E[M] \equiv \langle M \rangle = Np \tag{2}$$
$$V[M] \equiv \langle M^2 \rangle - \langle M \rangle^2 = Np(1-p) \tag{3}$$

で与えられる (以下 $E[M]$, $\langle M\rangle$ などの表記を適当に使い分ける)．二項サンプリングはすべてのノードを平等に確率 p でサンプルするが，二項分布は $N, Np, Np(1-p) \to \infty$ で正規分布に，$N \to \infty, p \to 0, Np =$ 一定でポアソン分布に近づくので，パラメータの設定次第で他のタイプのサンプリングに対応させることもできる．実際この論文では $p \to 0$ の極限でサンプルの次数分布を計算するので，ポアソンサンプリングの場合の次数分布に関する定理とみなすこともできる．確率 p はサンプリングの「エフォート」に対応するので，論文の概要にある「parsimonious=倹約的な」は p が小さく最小限のエフォートであることを表現している．

サンプルされた部分ネットの次数分布を $P^*(k)$ と書く．ここで，次数分布の**確率母関数** (probability-generating function: PGF) を考える．分布と母関数は完全に同等の情報を持っており，お互いをお互いから導くことができるので，ここではネットワーク全体の次数分布 $P(k)$ と部分ネット（サンプル）の次数分布 $P^*(k)$ を直接比較するのではなく，対応する確率母関数を計算してそれらが一致するかどうかを調べる．PGF の定義は

$$G(s) \equiv \sum_{i=0}^{\infty} P(i) s^i \qquad \text{(本文中の式番号 [2])} \tag{4}$$

で与えられ，これを微分することにより次数分布

$$P(k) \equiv \frac{1}{k!} \left. \frac{\mathrm{d}^k G(s)}{\mathrm{d}s^k} \right|_{s=0} \tag{5}$$

が得られる．

（注意）スケールフリーネットワークでは「孤児 (orphan) ノード」は考えないので，$P(0) = 0$ である．

フルネット $\mathcal{N}$ の中で次数 i をもつノードがサブネット $\mathcal{S}$ で次数 $k(\leq i)$ をもつ確率は二項分布 ${}_iC_k p^k (1-p)^{i-k}$ なので，サブネットの次数分布は

$$P^*(k) = \sum_{i \geq k}^{\infty} P(i) {}_iC_k p^k (1-p)^{i-k} \qquad [3] \tag{6}$$

で与えられる．これにより，サブネットの PGF は，

$$\begin{aligned} G^*(s) &\equiv \sum_{k=0}^{\infty} P^*(k) s^k = \sum_{k=0}^{\infty} \sum_{i \geq k}^{\infty} P(i) {}_iC_k p^k (1-p)^{i-k} s^k \\ &= \sum_{i \geq k}^{\infty} P(i) \sum_{k=0}^{\infty} {}_iC_k (ps)^k (1-p)^{i-k} = \sum_{i=0}^{\infty} P(i) \left[ps + (1-p)\right]^i \\ &= G(ps + (1-p)) = G(1 - p(1-s)) \qquad [4] \end{aligned} \tag{7}$$

２行目では付録 A にある二項展開の公式を，３行目では [2] 式を用いた．

（注意）明らかに規格化条件 $G^*(1) = G(1) = \sum_{i=0}^{\infty} P(i) = 1$ が成り立つ．

二項サンプリングだけでは部分ネット $\mathcal{S}$ の中には孤児ノードが含まれる．その確率は，

$$P^*(0) = \sum_{i=0}^{\infty} P(i)(1-p)^i = G(1-p) = G^*(0) \tag{8}$$

である．最初の等号で [3] 式，次の等号で [2] 式，最後の等号で [4] 式を使った．

二段サンプリング (subsampling)

さらに孤児ノードを除く二段サンプリングを行う．本文ではこのサンプリングの後のサブネットの次数分布も $P^*(k)$ と表記しているが，紛らわしいので，このメモでは二段サンプリングの後のサブネットの次数分布を $P^{**}(k)$，対応する確率母関数を $G^{**}(s)$ と表記することにする．Supporting Text には「二項二段抽出のもとで閉じた (closed under binomial subsampling) 分布関数族の集合」に関する定理の数学的な証明が書いてある．

まず，孤児ノードを除いた後の再規格化定数 C を求める．$C(1-P^*(0)) = 1$ より，

$$C = \frac{1}{1-P^*(0)} = \frac{1}{1-G(1-p)} \tag{9}$$

となる．最後の等式で (8) 式を使った．本文脚注には

$$\begin{aligned} \frac{1}{C} &= 1 - P^*(0) = 1 - \sum_{i=0}^{\infty} P(i)(1-p)^i = \sum_{i=0}^{\infty} P(i)\left[1-(1-p)^i\right] \\ &= 1 - G(1-p) \end{aligned} \tag{10}$$

という式も書いてある．二つ目の等式で [3] 式を，三つ目の等式では規格化条件 $1 = \sum_{i=0}^{\infty} P(i)$ を，最後の等式では [2] 式を使った．これにより，二項二段サンプリングの後の次数分布 $P^{**}(k)$ に対応する確率母関数 $G^{**}(s)$ が，

$$\begin{aligned} G^{**}(s) &\equiv \sum_{k=0}^{\infty} P^{**}(k)s^k = C\sum_{k=1}^{\infty} P^*(k)s^k \\ &= C\left\{\sum_{k=0}^{\infty} p^*(k)s^k - P^*(0)\right\} \\ &= \frac{G(1-p(1-s)) - G(1-p)}{1-G(1-p)} \qquad [5] \end{aligned} \tag{11}$$

のようにフルネットの確率母関数 $G(s)$ の関数として与えられる．最後の等式で (7) 式と (8) 式を使った．[5] 式は，明らかにフルネットの PGF（[2] 式）とサブネットの PGF（[4] 式もしくは [5] 式）が一般には異なること，すなわち，フルネットの次数分布とサブネットの次数分布が異なることを示している．

一般に，フルネットの次数分布とある次数分布の族の中のサブネットの次数分布が同型になるためには

$$G^*(s,\Omega) = G(s,\Omega') \qquad [6] \tag{12}$$

となる必要がある．ただし，Ω, Ω' は分布を特徴付けるパラメータである．[4] もしくは [5] が [2] と条件 [6] を満たすための必要十分条件は

$$G^*(s,\Omega) = G(1-(1-p)s,\Omega) = G(s,\Omega') \qquad [7] \tag{13}$$

となる（証明は Supporting Text 参照）．

正または負の二項分布に関しては [6] が満たされる．すなわち，あるネットワーク全体の次数分布が正または負の二項分布に従うなら，二項二段サンプリングされたサブネットの次数分布も正または負の二項分布に従う．もちろん平均次数はサンプリング確率 p の分だけ小さくなるが，異なる値のパラメータをもつ「同じ族」に属するという [6] 式の意味で「同じ分布に従う」．正または負の二項分布は，Erdös-Rényi 型のランダムグラフの次数分布 （もしくは古典的なランダムグラフもしくはポアソン分布）や指数分布を含む広いクラスの分布であることに注意．すなわち，スケールフリーネットワークを除く多くの一般的なネットワークにおいてはフルネットと部分ネットは一致する．

実際，二項サンプリングされたサブネットの確率母関数（[4] 式）に付録 B で求めた負の二項分布の確率母関数 (B.8) を代入してみると，

$$\begin{aligned} G^*(s) &= G(1-p(1-s)) = \left[1+\frac{m}{k}(1-(1-p(1-s)))\right]^{-k} \\ &= \left[1+\frac{mp}{k}(1-s)\right]^{-k} \end{aligned} \tag{14}$$

となり，p4222 脚注 †† 中の式が得られる．これはフルネットの PGF（(B.8) 式）と同型である．平均が p 倍されただけで，”clumping”パラメータ k は変わらない．二項二段サンプリングされたサブネットの確率母関数（[5] 式）に (B.8) を代入する場合も，[5] 式の $G(1-p(1-s))$ 以外の項は s を含まず，$P^{**}(k)$ を生成する時に s で微分すると消えてしまうため，本質的には $G^{**}(s)$ は $G^*(s)$ や $G(s)$ と同型であるとみなしうる（次数分布は再規格化定数 C 倍されるだけ）．

一方，p が十分小さい場合には，フルネットが指数 γ のスケールフリーネットワーク（$P(k)\ k^{-\gamma}$）に対する，サブネットの次数分布を解析的に評価することができる．$\gamma = 2$ に対しては，

$$P^{**}(1) \simeq \frac{\ln\left(\frac{1}{p}\right)}{1+\ln\left(\frac{\mathrm{e}}{p}\right)} \tag{15}$$

$$P^{**}(k>1) \simeq \frac{[const.]}{k(k-1)} \tag{16}$$

となり，サブネットの次数分布はフルネットの次数分布とは異なることがわかる．同様に，$\gamma = 3$ に対しても

$$P^{**}(k>2) \simeq \frac{[const.]}{k(k-1)(k-2)} \tag{17}$$

となり，やはりフルネットの次数分布とは一致しない．

図２は，$\gamma = 3, 2, 1.5$ のそれぞれに対して，フルネット，$p = 0.8$ および $p = 0.2$ の場合のサブネットの次数分布をグラフ化したものである．γ が大きいほど，また p が小さいほど，フルネットとサブネットの次数分布のずれが大きくなることがわかる．また，サブネットにおいては相対的に次数が少ないノードが増える．次数分布を両対数で見た場合，フルネットは定義上直線のグラフになるが，サブネットは，凹（下に凸）(concave) 型のカーブになる．著者らは，実在するサンプルされたネットワークの次数分布は凸（上に凸）(convex) 型のカーブになっていることが多く（ハブは相対的に少ない），そのため実在のネットワークを完全にサンプルしたフルネットはスケールフリーからはずれる可能性を指摘している．一般的なファットテール的な次数分布に対してもスケールフリーと分類してしまっている可能性がある．

結論

著者らは，生物や他の複雑ネットワークの役割を理解するためには，ベキ的なテール（ハブ）だけでなく，ネットワーク全体を見る必要があると指摘している．特に，次数が少ないノードほどサンプリングの影響を強く受けることには注意を払うべきであろう．次数が小さいノードは，pが小さいと二項二段サンプリングの後には孤児ノードになりやすく，また，pは「サンプリング・エフォート」にも対応するから，エフォートの少ない小サンプルからフルネットのスケールフリー性を推定する場合には特に注意が必要であろう．

感想

より現実的な状況に対応させるためには,「有限母集団」（N 有限）に対する「サンプリング定理」が必要になるだろう．また,「最も倹約的な ($p \ll 1$)」サンプリングだけでなく，有限「エフォート」の p に対する厳密な評価もできれば有用であろう．さらに，母集団（フルネット）の分布が陽に与えられておらず，Albert-Barabási 的なダイナミクスに対して，フルネットとサブネットの分布を同時に求める数学的方法の研究も興味深い．最近の群集生態学における「中立理論」は，その（数少ない）例のひとつであり,「死亡（ノードの確率的除去)」があるような,「成長・死亡する有限サイズネットワーク」の次数分布については，中立理論が応用できるかも知れない．

付録

論文には書かれていないが，以下に補足のための付録を付ける．

A　二項分布の期待値と分散

期待値や分散などの分布のモーメントを計算するときには，モーメント母関数を求めると便利である．二項分布のモーメント母関数 $K(t)$ は,

$$\begin{aligned} K(t) &\equiv \langle \mathrm{e}^{tM} \rangle = \sum_{M=0}^{N} \mathrm{e}^{tM} \mathrm{Bi}(M) = \sum_{M=0}^{N} \mathrm{e}^{tM} {}_N C_M p^M (1-p)^{N-M} \\ &= \sum_{M=0}^{N} {}_N C_M (p\mathrm{e}^t)^M (1-p)^{N-M} \\ &= \left[p\mathrm{e}^t + (1-p) \right]^N \end{aligned} \tag{A.1}$$

となる．最後の等号では二項展開の公式 $(a+b)^N = \sum_{M=0}^{N} {}_N C_M a^M b^{N-M}$ を使った．これにより機械的に期待値と分散が

$$E[M] \equiv \left. \frac{\mathrm{d}K}{\mathrm{d}t} \right|_{t=0} = \left. Np\mathrm{e}^t \left[p\mathrm{e}^t + (1-p) \right]^{N-1} \right|_{t=0} = Np \tag{A.2}$$

$$E[M^2] \equiv \left. \frac{\mathrm{d}^2 K}{\mathrm{d}t^2} \right|_{t=0} = (\cdots) = Np + N(N-1)p^2 \tag{A.3}$$

$$V[M] \equiv E[M^2] - E[M]^2 = Np + N(N-1)p^2 - (Np)^2 = Np(1-p) \tag{A.4}$$

と求まる．

B　負の二項分布の確率母関数

負の二項分布 (negative binomial distribution)

$$\mathrm{NB}(r;p,k) \equiv \left(\frac{r+k-1}{r}\right)(1-p)^k p^r = (-1)^r \left(\frac{-k}{r}\right) p^r (1-p)^k \tag{B.1}$$

は,「成功確率 p の事象が r 回成功するまでに k 回失敗したというベルヌーイ試行が起こる確率」である．負の二項分布に関しては，他にも異なる定義がいくつかあることに注意．また，本文 p4222 脚注 †† の定義では，k は変数ではなく"clumpiness"パラメータで，r が変数である．すなわち，$\mathrm{NB}(r;p,k)$ を次数分布とみなすときには，r が次数に対応する．この点，論文のノーテーションがややミスリーディングなので注意．(B.1) の二つ目の等号では

$$\begin{aligned}
\left(\frac{r+k-1}{r}\right) &= \frac{(r+k-1)!}{r!(k-1)!} = \frac{(r+k-1)(r+k-2)\cdots(k)}{r!} \\
&= (-1)^r \frac{(-k)(-k-1)(-k-2)\cdots(-k-(r-1))}{r!} \\
&= (-1)^r \frac{(-k)(-k-1)(-k-2)\cdots(-k-(r-1))}{r!}\frac{(-k-r)!}{(-k-r)!} \\
&= (-1)^r \frac{(-k)!}{r!(-k-r)!} \\
&= (-1)^r \left(\frac{-k}{r}\right)
\end{aligned} \tag{B.2}$$

と書けることを用いた．

まず，負の二項分布のモーメント母関数を求め，期待値と分散を求める．モーメント母関数は

$$\begin{aligned}
K'(t) &\equiv \left\langle \mathrm{e}^{tK} \right\rangle = \sum_{r=0}^{\infty} \mathrm{e}^{tr} (-1) \left(\frac{-k}{r}\right) p^r (1-p)^k \\
&= (1-p)^k \sum_{r=0}^{\infty} \left(\frac{-k}{r}\right) 1^r (-\mathrm{e}^t p)^r \\
&= (1-p)^k \sum_{r=0}^{\infty} \left(\frac{-k}{r}\right) 1^{-k-r} (-\mathrm{e}^t p)^r \\
&= (1-p)^k (1-\mathrm{e}^t p)^{-k}
\end{aligned} \tag{B.3}$$

となる．最後の等号で二項展開の公式を用いた．これを用いて，期待値，２乗平均，分散は

$$m \equiv E[r] \equiv \langle r \rangle = \left.\frac{\mathrm{d}K}{\mathrm{d}t}\right|_{t=0} = (\cdots) = \frac{kp}{1-p} \tag{B.4}$$

$$E[r^2] \equiv \langle r^2 \rangle = \left.\frac{\mathrm{d}^2K}{\mathrm{d}t^2}\right|_{t=0} = (\cdots) = \frac{k^2p^2+kp}{(1-p)^2} \tag{B.5}$$

$$\sigma^2 = V[r] = E[r^2] - E[r]^2 = \frac{kp}{(1-p)^2} \tag{B.6}$$

となる．これらを用いて p4222 脚注 †† の中の式

$$\sigma^2/m^2 = \frac{kp}{(1-p)^2}\left(\frac{1-p}{kp}\right)^2 = \frac{1}{kp} = \frac{1-p}{kp} + \frac{p}{kp} = 1/m + 1/k \tag{B.7}$$

が得られる．

また，負の二項分布の PGF は，上記の期待値 m を用いて，

$$\begin{aligned}
G(s) &= \sum_{r=0}^{\infty} \mathrm{NB}(r;p,k)s^r = \sum_{r=0}^{\infty} s^r(-1)^r\binom{-k}{r}p^r(1-p)^k \\
&= (1-p)^k\sum_{r=0}^{\infty}\binom{-k}{r}1^{-k-r}(-sp)^r \\
&= (1-p)^k(1-sp)^{-k} \\
&= \left(\frac{1-ps}{1-p}\right)^{-k} \\
&= \left(\frac{1}{1-p} - \frac{kp}{1-p}\frac{1}{k}s\right)^{-k} \\
&= \left(1 + \frac{p}{1-p} - \frac{m}{k}s\right)^{-k} \\
&= \left(1 + \frac{kp}{1-p}\frac{1}{k} - \frac{m}{k}s\right)^{-k} \\
&= \left[1 + \frac{m}{k}(1-s)\right]^{-k}
\end{aligned} \tag{B.8}$$

と表される．これが p4222 脚注 †† の中にでてくる．

参考文献

[1] R. M. May. Will a large complex system be stable? *Nature*, Vol. 238, pp. 413–414, 1972.

[2] R. M. May. Simple mathematical models with very complicated dynamics. *Nature*, Vol. 261, pp. 459–467, 1976.

[3] R. M. Anderson and R. M. May (eds). *Population Biology of Infectious Diseases*. Springer-Verlag, 1982.

[4] M. A. Nowak and R. M. May. *Virus Dynamics: the Mathematical Foundations of Immunology and Virology*. Oxford University Press, 2000.