

# 効率とシステムサイズの関係に関する確率モデルによる考察

早稲田大学先進理工学部物理学科 4 年 藤本將太郎 (1Y11A045-8)

2015 年 1 月 30 日

## 目次

1	導入	3
2	モデルの説明	4
3	各モデルにおける結果と考察	7
3.1	$X_i$ を等確率で選び、その中から $x$ が一様に選ばれる場合 . . . . .	7
3.2	$X_i$ を一つ前の $X_j$ によって決まる確率で選び、その中から点 $x$ が一様に選ばれる場合 . . . . .	10
4	結論	13
5	参考文献	13

## 1 導入

私達は普段の身近な生活の中で、サイズが変わるとその中の質や状態が変わってしまうと感じるものを見ることができる。それは会話とその人数の間の関係にもあるかもしれないし、共同作業を行うときのチームメンバーの人数かもしれない。企業の大きさとその中の人間の従事度も異なるかもしれない。会議にやたら多くの人が参加していても、ほとんどの人はその会議を無意味なものだと見なすことになるだろう。世の中のおよそほとんどのスポーツやゲームなどは、参加する人数が予め決められていることが多い。よく練られたものであれば、この決められた人数より多くても少なくとも、本来の楽しみを得ることはできないだろう。麻雀は4人でするから楽しめるのであり、会議は100人ではなかなか進まないものである。

このような現象はこれまで見てきたように大変身近で、私達にとって実感をもって受け入れられることなので、ほとんどの場合不文律にこの法則は受け入れられていると言って良いだろう。

また、相乗効果という点に着目すれば、一人で何かするよりも、多くの人を交えて行ったほうが効率が上がるといったものも多く存在していることだろう。しかしながら、そこでも多すぎはまた別の問題を産むなどして、多すぎることも効率を下げる要因になることもあるのである。

こういったことを私達は無自覚の内に理解しているので、何かを大人数で共同して行うときには、それをさらにいくつかのグループ、班、部署、そういったより小さい単位に分割する必要があると感じる。もちろん、このような現象はよく知られたことであるので、どれほどのシステムサイズであれば最大の効率が導き出せるか、ということは、昔から考えるべきことであり、最適人数に関する調査・研究が行われたり、詳細な研究はなされていなくても各分野での情報が蓄積されたりしている。例えば教育の分野などではグループ学習や少人数授業の有効性が議論されることも多い。

さて、これまでに述べたような系のシステムサイズとその効率の間の関係は、何人も人間を要素とした系でなくとも考えることが出来る。例えば、動物の体重と代謝率の間の関係も、同じ現象であると思なすことができるかもしれない。動物学等の分野で知られている法則として、体重とその他の観測量の間の関係としてアロメトリー則と呼ばれるものがある。その一つが体重と代謝率の間の関係であり、特にこの関係のことを Kleiber 則と呼ぶ。これは様々なスケールの対象について調べられていて、代謝率  $E$  と体重  $M$  の間の関係は  $E \sim M^b$  のようにスケールされる。この指数については、データの取り方などによって異なるため、いくつか説があり、指数の大きさは  $2/3$  とする意見と  $3/4$  とする意見とがあり、また両辺の対数を取った関係式で右辺の  $\log M$  の2乗の項を考慮すればよりよくフィッティングできる、といったものもある。ここで重要視したいのは、指数の大きさがどちらか、ということではなく、その指数は1よりも小さい値をとる、ということである。これまで挙げてきたシステムサイズと効率の関係についてと同じように議論する余地があることが分かる。

この指数が生物種や成長過程における細胞組成の変化が本質的ではなく、その時点でのシステムのサイズが影響していることの一つの例として、ホヤの群体サイズと代謝率の間の関係も、同様に1乗より小さい指数でスケールされることが確かめられている。このホヤは無性生殖によって自分と全く同じ個体を増やして群体を形成し、群体同士のつながりはネットワーク上に張り巡らされた血管のつながりである。この群体の構成個数は実験者が物理的に切り離すことによって容易に変更でき、それによってサイズごとの代謝率を調べることができる。さらに、このホヤにはある時間周期ごとに一斉に分裂・増殖を行う“takeover”と呼ばれる現象が見られ、このときすべての個体間の血管つながりが切れ、このときの代謝率は同じ個体数の群体の代謝率よりも大きくなることが示されている。二つ目の例の場合は、これから増殖を行うために代謝率が増加したのだと見ることもできるのであるが、一つ目のサイズと代謝率の間の関係は、そのように簡単に説明されるものではな

い。これまでの Kleiber 則の説明としては、体温の維持と表面積との関係から  $2/3$  を導くものや、血管構造が大動脈から毛細血管に至るまでがフラクタル的であることに着目し、血液の流速と代謝率の間の関係を仮定して  $3/4$  を導くものがあった。しかしはじめの理論では変温動物や単細胞においてもこの関係が成り立つことのうまい説明ができない。また、二つ目の理論についても、より物理的な描像なため説得力はあるが、ホヤの場合、血管構造、特に血管の太さや長さに関して自己相似的とは言えず、また、その理論の中で仮定される”巨視的プールから微視的構造に向かう流れ”というものも見られないため、この理論では説明することはできない。したがって、引用元の記事の中では一つの仮説として以下のようなものが提唱されている。

『べき乗のサイズ効果は個体同士の局所的な相互作用によって生じ、takeover 中には個体間の連絡が切れるため、サイズ効果が見られなくなる。同じユニットが局所的な相互作用をもつ系は、自己組織化臨界状態にある可能性があり、臨界状態とは相互作用に効果がべき関数で記載されるものである。ホヤを含め、動物は自己組織化臨界状態にあるのではないか。』

ここまで述べて来たような、人間の作るシステムの中での最適人数の議論や、生物のアロメトリー則などは、統一した一つの理論として書ける可能性がある。このとき、それまで各モデルの中で重要だと思われてきた要素は、統一された理論の中でのある物理量に対応させることによって、どのモデルでも同じように書けるかもしれない。こうしたことを考えていくうちに、より抽象的にこの問題を考えることには何か意義があるのではという思いを抱くようになった。

この論文では、具体的なイメージとしては例の中で挙げられた会議を題材に議論を進めることにした。会議の進行プロセスを元に確率モデルを作成し、その解析を行って、その結果をまとめた。

## 2 モデルの説明

まずはじめに、作成した確率モデルに関しての説明と、その際追加で説明が必要と思われた場合にはその背景についても述べていくことにする。

会議について考える上で、システムサイズは会議に参加する人の人数であるとする。したがって、これから考えていくべき関係はこの人数と会議の”質”、成果との間の関係である。しかしながら、この”質”がどういった量として測れるのかは現段階で不明であるため、モデルを立てた際に得られる一般的な物理量から、質にあたるものを予測していく、というアプローチを取ることにする。

一口に会議と言っても世の中には様々な種類の会議が存在する。会社やグループの中で行われる会議の中には、すべての部署が一同に会し、その中でそれぞれの進捗状況など、情報を共有するための報告会議や、企画のためのアイデア出しなど、何か与えられた議題に沿ってそれぞれが自由に発言するタイプの会議もある。また、既に決定している内容に関して、それに関連のある人を集めて一度に口頭で説明するタイプの報告会も、一般には会議と呼ばれている。また、会議の進行方法もいくつかあり、会話と同じようにそれぞれの人が思い思いに発言する場合や、ファシリテーターと呼ばれる進行役が質問によってそれぞれの人の意見を引き出したり、意見をまとめたりして、会議の流れをコントロールする場合もある。このモデルで扱うのは、アイデア出しのようにそれぞれが自由に発言し、その意見がつながることによってよりよい案を得ることができる、という種類の会議であり、ファシリテーターのように全体の流れをコントロールできる独立した存在はいないものとする。

以上のような設定を考えた上で、これを確率論として扱うために、会議の中の言葉を抽象的なモデルで表すことにする。まず、「意見」とはまだ会議の中で発言されていない、それぞれの人がもつ考えをあらわすこととする。すべての意見が  $a$  個の異なる要素に分解でき、それぞれの要素は 1 つの実数値をもっているとする

と、ある一つの意見  $x$  というのは、 $\Omega \subset \mathbb{R}^a$  上の一つの元

$$x = (x_1, x_2, \dots, x_a) \in \Omega$$

とすることができる。

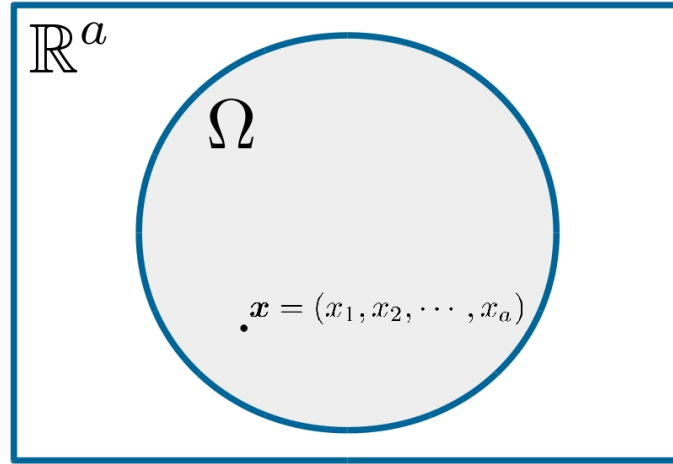


図1  $\mathbb{R}^a$  上の空間  $\Omega$  と、その元である点  $x$

また、この意見間の距離を定義すればこれらの点を元とする距離空間を定義することができる。距離の入れ方としてはユークリッド距離

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_a - y_a)^2}$$

を考慮することとする。

次に、「参加者」について考える。参加者は会議の中で自分の中にある意見を発言していくのであり、参加者それ自体は意見の集まりである。すなわち意見空間  $\Omega$  上の部分空間  $X_i$  が参加者  $i$  を表すとする。 $x_{ik}$  が人  $i$  のラベル  $k$  の意見であるとする、人  $i$  が  $s_i$  個の意見を持っているとき、

$$X_i = \{x_{i0}, x_{i1}, \dots, x_{is_i}\}.$$

一般に

$$X_i \cap X_j = \emptyset \quad \text{when} \quad i \neq j$$

である必然性はないが、このモデルでは上式を採用することとする。したがってモデルの中で扱う意見空間  $\Omega_0$  は、参加者が  $N$  人集まったとき、 $X_i$  の直和集合

$$\Omega_0 = \bigcup_{i=1}^N X_i \quad (1)$$

$$= \sum_{i=1}^N X_i \quad (2)$$

となる。

モデルの作成に際し、会議の進行に関していくつかの仮定をおいて考えていくことにする。これらの仮定が本質的であるような場合にはまた別のモデル化を考える必要があるが、ひとまず以下の仮定を受け入れることにして、その上でモデルを作成することにした。

仮定:

- 一度に発言できる人は一人まで
- 一人当たりの発言時間は参加人数に依らない

このように仮定すると、会議という過程を、1つの時間発展するネットワークの生成過程であると見なすことができる。すなわちある参加者に必ず属する点をルールにしたがって各時刻ごとに選択していき、その時意見間の間にも与えられた別のルールによってエッジを結んでいくとする。このようにすると、終了時刻  $K$  が経過した時には、意見間のネットワークと、参加者間のネットワークを得ることができる。このネットワークの特性の中に、会議の質となりうる量があるとして、この描像のもとに意見  $x$  や参加者  $X_i$  の選択に関してのルールを与えていくことにする。

以降の議論では、特に必要のない限り、上で定めた記号と言葉を使い、「会議、意見、参加者」といった言葉は使わないことにする。

点  $x$  の選び方としては、

1. 集合  $X_i$  を選び、その中から  $x$  を選ぶ方法
2. 点  $x$  を選び、その点の情報から、 $X_i$  を決める方法

の二つがある。

それぞれの場合について、 $x$  と  $X$  を選ぶ確率が互いに独立な場合と、そうでない場合にさらに分けて考えることができる。

したがって、考えるべきは

- $X_i$  の独立な選び方
- $x$  の独立な選び方
- $X_i$  が選ばれたときの  $x$  の選び方
- $x$  が選ばれたときの  $X_i$  の選び方

の4つである。

### 3 各モデルにおける結果と考察

この章では、2章で設定したモデル化の方法を用いて、 $x$  と  $X_i$  の選択に関するルールと、 $x$  間のエッジの結び方に関するルールを定め、そのルールに従ったときにはどのようなネットワークの特徴量が得られるかについて、それぞれ見ていくことにする。

#### 3.1 $X_i$ を等確率で選び、その中から $x$ が一様に選ばれる場合

この設定は、はじめに  $N$  個の集合  $X$  から 1 つの  $X_i$  を選択し、点  $x$  はそれぞれの  $X_i$  について  $\Omega = [0, 1]$  に一様に分布しているとする。このとき、 $X_i$  の数  $N$  が変化しても、 $x$  についての選び方は同じであるから、 $x$  の作るネットワークだけでは、本来考えたい  $N$  との関係は見ることはできない。しかしながら、選ばれた点のネットワークに関する性質を調べるには良い練習となる。

$X_i$  から  $k$  番目に点を選ばれ、その後時刻  $k+1$  に  $X_j$  から点を選ばれる確率は、 $X_i$  の数  $N$  として

$$p_i(k, j) = p(N) = \frac{1}{N}$$

のように、直前の履歴に依存せず、等確率である場合を考えている。

このとき、 $x$  は  $[0, 1]$  の間の値を一様な確率で取るとしていたので、そうして得られた確率変数  $x$  に関して、確率密度関数  $f(x)$  は

$$f(x) = 1 \quad (0 \leq x \leq 1) \quad \text{otherwise } 0$$

であり、累積分布関数  $F(x)$  は

$$F(x) = x \quad (0 \leq x \leq 1)$$

である。

時間発展によって点選ばれていくごとに、その点と距離  $r$  より近い位置にある点との間にすべてエッジを張っていくことにする。すなわち、この問題はある閾値  $r$  ( $0 < r \leq 1$ ) を定めた時に、時刻  $k$  で選ばれた点  $x_k$  を中心とする領域  $[x^k - r, x^k + r]$  の中に入るそれまでに点の数はいくつか、という問題に帰着される。

すなわち図 2 のような状況を考えることになる。

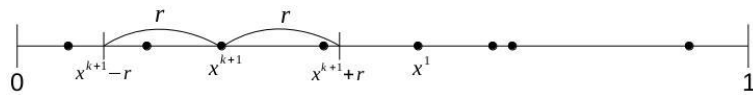


図 2  $[0, 1]$  の数直線上で閾値  $r$  で定められる領域

一様な確率で  $[0, 1]$  の間の数が選ばれるとき、その確率変数が  $[\max(0, x - r), \min(x + r, 1)]$  の範囲に入っている確率は、確率密度関数を用いて、

$$p(\max(0, x-r), \min(x+r, 1)) = \int_{\max(0, x-r)}^{\min(x+r, 1)} 1dx \quad (3)$$

$$= [x]_{\max(0, x-r)}^{\min(x+r, 1)} \quad (4)$$

$$(5)$$

よって

$$p(x, r) = \begin{cases} x+r & 0 \leq x < \min(r, 1-r) \\ p(r) = \min(2r, 1) & \min(r, 1-r) \leq x \leq \max(1-r, r) \\ 1-x+r & \max(1-r, r) < x \leq 1 \end{cases} \quad (6)$$

を得る。

これはグラフにすると、図3のようになる。

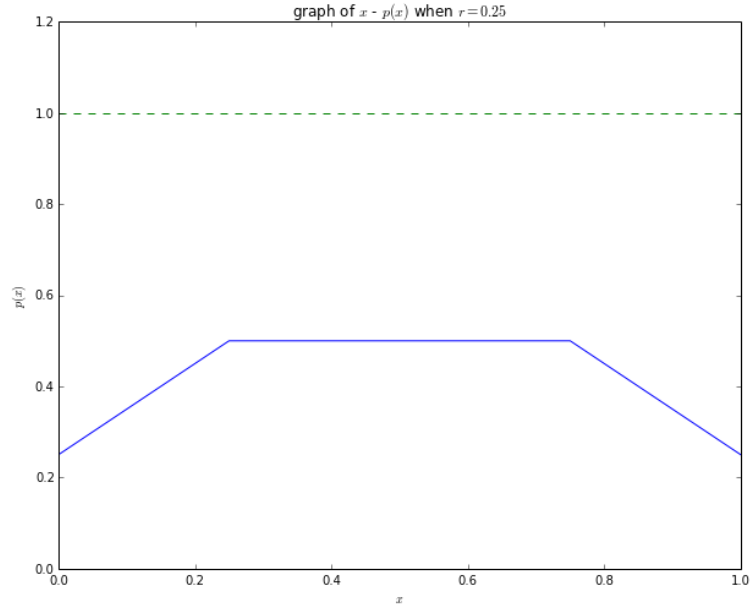


図3  $r = 0.25$  のとき、それぞれの位置  $x$  において他の点を1つ見出す確率  $p(x)$

$k+1$  番目に点  $x_{k+1}$  が選ばれたとき、 $k$  番目までに選ばれた点のうち  $y$  個の点が領域  $[\max(0, x-r), \min(x+r, 1)]$  の中に存在する確率は、

$${}_k C_y p(r)^y p(r)^{k-y} \quad (7)$$

で表せる。ここで注意すべき点は、 $p(x_k, r)$  は  $x_k$  によって変わるものであったから、 $x$  に対して期待値を取ったものを考えなければならないことである。 $p(r)$  はそのようにして得られた期待値であることを意味している。

$p(r)$  を求めると、



$0 < r \leq 0.5$  のとき

$$p(r) = E(p(x, r)) = \int_0^r x + r dx + \int_r^{1-r} 2r dx + \int_{1-r}^1 1 - x + r dx \quad (8)$$

$$= \left[ \frac{x^2}{2} + rx \right]_0^r + [2rx]_r^{1-r} + \left[ x - \frac{x^2}{2} + rx \right]_{1-r}^1 \quad (9)$$

$$= -r^2 + 2r \quad (10)$$

$0.5 < r \leq 1$  のときも同様にして

$$p(r) = E(p(x, r)) = -r^2 + 2r$$

である。

また、式 (7) は

$$P[X = y] = {}_k C_y p(r)^y (1 - p(r))^{k-y}$$

のように書けば明らかなように、確率変数  $X$  に対するパラメータ  $k, p$  の二項分布  $B(k, p)$  を表している。  
この時、確率変数  $X$  に対する期待値と分散は、

$$E(X) = kp(r)$$

$$V(X) = kp(r)(1 - p(r))$$

である。

時刻  $k$  が大きい時、具体的には

$$kp(r) > 5,$$

$$kp(r)(1 - p(r)) > 5$$

をみたす  $k$  のとき、二項分布は正規分布に近似できるので、期待値は中央値とほぼ等しくなる。すなわち、 $y = kp(r)$  より多くの点を見出す確率  $P[X \leq y]$  は、どんな時刻  $k$  においても  $1/2$  となる。

すなわち、時刻の依存性はないため、 $x_{k+1}$  のまわりの領域に  $y$  個以上の他の点が存在する場合を 1、存在しない場合を 0 とすれば、これらの起こる確率は、それぞれ  $1/2$  であり、 $k$  回目に  $y$  個以上のエッジが張られる確率は、 $p = 1/2$  の幾何分布に従う：

$$P[X = k] = \left( \frac{1}{2} \right)^k$$

このときの期待値と分散は

$$E(X) = \frac{1}{p} = 2$$

$$V(X) = \frac{1-p}{p^2} = 2$$

すなわち、これまでのような場合を考えたとしても、平均として 2 回で  $y$  個以上のエッジが張られ、ほとんどの試行は 13 回でその値に到達することが分かる。しかし、これは先程述べたような二項分布の正規分布への近似ができない領域であることに注意が必要である。

### 3.2 $X_i$ を一つ前の $X_j$ によって決まる確率で選び、その中から点 $x$ が一様に選ばれる場合

次に考えるのは、 $X_i$  の選び方が、一つ前の  $X$  による確率で決定するような場合である。この確率を定めるにあたり、会議として自然と思われる  $X_i$  間に距離が定義でき、その距離にしたがって確率が決まるような問題を考えることとした。このとき距離の計算に用いることができるパラメータの数を  $b$  とし、距離として  $b$  次元ユークリッド距離を考えることにする。

すなわち  $b$  個のパラメータを要素とする元からなる空間  $X$  があったとき、距離関数  $d: X \times X \rightarrow \mathbb{R}$  が

$$d(x, y) = \sqrt{\sum_{i=1}^b (x_i - y_i)^2} \quad , \quad x, y \in X$$

と書けることを意味する。実際の場合には、各データ同士の相関を考慮に入れたマハラノビス距離などのほうが適当な場合もあるかもしれないが、まずはイメージしやすいということでユークリッド距離を考えた。

以下では、記述を簡単にするため、 $X_i$  と  $X_j$  の間の距離を  $d_{ij}$  と書くことにする。

時刻  $k$  に  $X_i$  が選ばれ、その後時刻  $k+1$  に  $X_j$  から点  $x_{k+1}^j$  が選ばれる確率  $p_k(i, j)$  は、距離  $d_{ij}$  の関数として、次のようにできる。

$$p_k(i, j) = \frac{g_k(d_{ij})}{\sum_j g_k(d_{ij})}$$

ここでの関数  $g$  の選び方によって、距離の大きさがどのように確率に重みを持たせるかということが決定される。一般に  $g$  は時刻  $k$  によって変化してもいいので、添字  $k$  をつけて時刻  $k$  における関数であることを表した。

単純な例として  $g_k(d) = \text{const.}$ ,  $\forall k, d \in \mathbb{R}^1$  とすると、距離に依らず  $X_i$  が選ばれるわけなので、3.1 節の  $X_i$  の選び方と同じである。

$g(d)$  は  $[0, +\infty]$  で定義される非負の実関数であればよい。

ex)

$$g(d) = \frac{1}{d+1}$$

$$g(d) = e^{-d}$$

$$g(d) = \begin{cases} c & (0 \leq d \leq 1/c) \\ 1/d & (d > 1/c) \end{cases} \quad , \quad c > 0$$

しかし、ここで注意すべき点として、どの  $X_i$  が選ばれたとしても、3.1 で考えたように、どの  $X_i$  も  $[0, 1]$  から一様に点  $x$  を取るから、結局点  $x$  について見たときの試行は同様のことをしており、どの  $X_i$  が選ばれるかは本質的な問題にはならないことが分かる。

これまでの設定を用いて数値シミュレーションを行った結果を図4~6に示す。シミュレーションでは、確率を決める距離の関数  $g(d)$  として

$$g(d) = e^{-d}$$

を採用し、 $x$  の次元  $a = 2$ 、 $X_i$  の数  $N = 6$  と設定した。

図4の中の青い丸が  $X_i$  の位置を表しており、それぞれの丸の大きさは、一回の試行において選択された頻度を表したものととなっている。円の間に張られた線分とそこに記された数字の組は、1番目の数字をラベルとしてもつ  $X_i$  のあとに2番目の数字をラベルとしてもつ  $X_j$  が選ばれたことを意味している。図5で青色の曲線で表されているのは、時刻  $k$  に選ばれた点  $x$  のまわりの  $r$  で決まる領域の中に入った、それまでに選択された点の数である。また、このグラフで緑色の直線として表されているものは、3.1のように計算で求めた値であり、

$$l = (-r^2 + 2r)k$$

であった。同じようにして偏差についても計算ができており、 $V(l) = \sqrt{(-r^2 + 2r)(1 + r^2 - 2r)k}$  である。これはグラフにおいて緑色の領域として描かれている。図6は、時刻  $k$  までに張られたエッジの数の総和を表しており、この中の緑色の曲線も、計算で求めることのできるものであった。

$$L = \frac{1}{2}(-r^2 + 2r)k^2$$

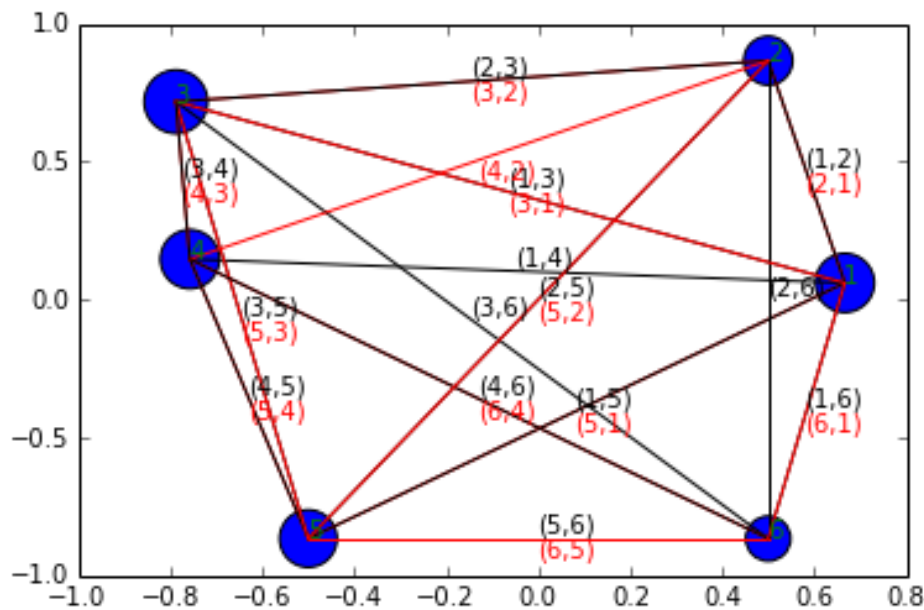


図4  $X_i$  の選択された頻度と  $X_i$  間ネットワーク

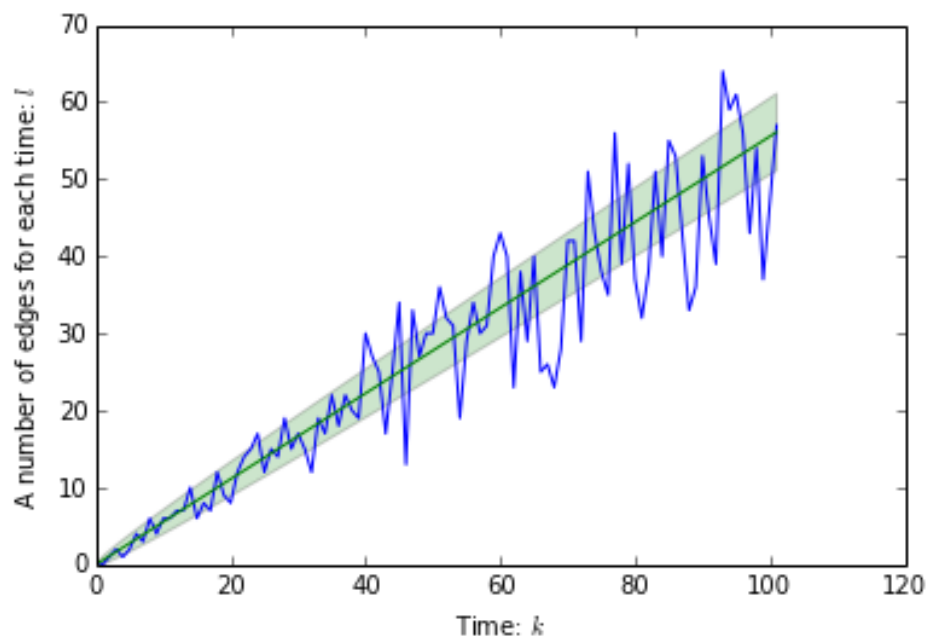


図5 時刻  $k$  とその時張られたエッジの本数  $l$  との関係

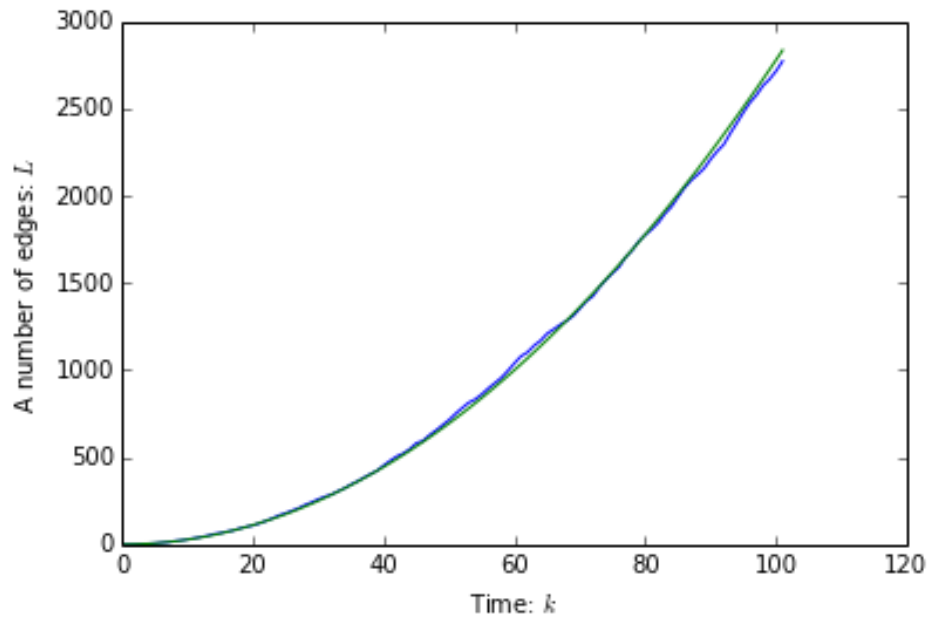


図6 時刻  $k$  までに張られたエッジの数の総和  $L$

4 結論

5 参考文献