千葉大学大学院工学研究科

修　士　論　文

複数のウォーカーを用いたランダムウォークサンプリングの性能評価

平成 29 年 2 月提出

指導教員：塩田　茂雄　教授

建築・都市科学専攻　都市環境システムコース

15TM0317　亀村　昂汰

Abstract

In this paper, we consider random walk sampling using multiple walkers combined. In random walk sampling, nodes in the network are not necessarily sampled with equal probability and depend on how the walker moves, so that the nodes to be sampled are biased. In this paper, we propose a method to estimate the true node information by grasping sampling bias in advance, statistically removing sampling deviations from collected data after node information collection. In the proposed method, it is possible, for example, to intentionally sample a specific node at a high frequency to collect information, then eliminate sampling bias and obtain true node information. In addition to showing the theoretical backing of the proposed method, we report the result of verifying the effectiveness by simulation.

概要

近年、Twitter や Facebook に代表されるオンラインソーシャルネットワークは爆発的な成長を遂げている。このようなネットワークは大規模かつ複雑なグラフデータとして知られている。

膨大なネットワークの特性、構造を把握するためには、ネットワークを構成するノードの情報を収集・分析する必要があるが、全てのノードの情報を収集することはコストや時間がかかり難しい。それゆえ、ランダムに選んだ一部のノードの情報を収集・分析する方法がしばしば用いられる。

本論文では、ランダムに一部のノードを選ぶ方法として、ウォーカーがネットワーク内をランダムに動き回りながら訪問ノードの情報を収集する「ランダムウォークサンプリング」のうち、複数のウォーカーを組み合わせて用いる手法について考察する。ランダムウォークサンプリングでは、ネットワーク内のノードは必ずしも等確率にサンプリングされず、ウォーカーの動き方に依存するため、サンプリングされるノードに偏りが生じる。本論文では、サンプリングの偏りを事前に把握し、ノード情報収集後に、収集データからサンプリングの偏りを統計的に除去して、真のノード情報を推定する手法を提案する。提案手法では、例えば意図的に特定のノードを高頻度でサンプリングして情報を収集し、その後、サンプリングの偏りを除去して真のノード情報を得ることなどが可能になる。本論文では、提案手法の理論的な裏づけを示すとともに、有効性をシミュレーションにより検証した結果について報告する。

目次

第 1 章

序論

1.1　研究背景・研究目的

近年，Twitter や Facebook に代表されるオンラインソーシャルネットワークは爆発的な成長を遂げており，オンラインソーシャルネットワークに参加しているユーザの特徴や行動特性などを踏まえて，オンラインソーシャルネットワーク上の様々な情報を分析し，そこから有意な知見を得る手法に関心が高まっている．

オンラインソーシャルネットワークの分析のためには，参加ユーザの特徴を知る必要があるが，ネットワークの運営者を除き，参加ユーザの全データを入手することは事実上不可能である．そのため現実には，一定の確率で抽出した一部の参加ユーザ（例えば 10%のユーザ）の特徴を分析することで，全参加ユーザの分析に代えることになる．しかし，ユーザ集合の全貌が不明な状態で，ユーザの一部を偏りなく抽出することも難しい．

ユーザをランダムに抽出し，その情報を収集する方法として，オンラインソーシャルネットワーク上をある種のプログラム（以下，ウォーカー）が動き回りながら，訪問したユーザの情報を収集する「ランダムウォークサンプリング」が提案されている．残念ながら，ランダムウォークサンプリングで抽出されるユーザには偏りがある．ユーザを「ノード」，ユーザ間のつながりを「リンク」として，オンラインソーシャルネットワークを抽象的な「グラフ」として捉えたとき，ウォーカーが隣接ノードの一つを等確率に選んで遷移することを繰り返すと，結果的にウォーカーは次数に比例する確率でノードを訪問し，ウォーカーが収集する情報も，高次数ノード（ユーザ）の情報に偏ることが知られる[1].

ウォーカーの高次数ノードへの訪問の偏りを取り除くため，ウォーカーの隣接ノードへの遷移確率を隣接ノードの次数に応じて調整し，均一なサンプリングを実現する手法（Metropolis-Hastings アルゴリズム）が提案されている[2]．一方，高次数ノードに偏って収集された情報に含まれるバイアスを事後処理で取り除く方法も提案されている．さらに両者を一つの枠組みで取り扱えるように拡張し，収集したい情報の種類に合わせて，意図的に特定のノード群（例えば，低次数ノード群）を高頻度に訪問して情報を収集し，収集後に情報に含まれているバイアスを除去して偏りのない特徴量を得る手法も提案されている[5]．しかし，これらは基本的に単一ウォーカー，もしくは複数のウォーカーを独立に動かす場合にのみ適用可能である．

本研究では，複数のウォーカーを組み合わせて用いるランダムウォークサンプリングについて考察する．複数のウォーカーを連携させて動かし，独立に動かす場合よりもサンプリングの精度を高める「Frontier Sampling」と呼ばれるアルゴリズムが提案されている[3]．本研究では，単一ウォーカー用に提案された手法[5]を，Frontier Sampling に適用できるように拡張する．これにより，複数のウォーカーを効率的に用いながら，分析の目的に合わせて，意図的に特定のノード群を高頻度に訪問して情報を収集し，事後に情報に含まれるバイアスを除去して真の特徴量を得ることが可能となる．本研究では，提案手法の有効性をシミュレーションにより検証した結果について報告する．

1.2　論文の構成

第 1 章　序論

本研究の概要について述べる.

第 2 章　ネットワーク

ネットワークに関しての解説をする.

第 3 章　ランダムウォークサンプリング

提案手法のランダムウォークサンプリングと, バイアス除去法の概要について述べる.

第 4 章　複数のウォーカーによるランダムウォークサンプリング

提案手法の複数のウォーカーによるランダムウォークサンプリングの概要について述べる.

第 5 章　シミュレーション実験

ネットワークのトポロジーデータを用いて, コンピュータ上に仮想的に構築したネットワークデータに, 提案手法を適用し有効性を評価する.

第 6 章　結論

本研究のまとめと今後の課題について述べる.

第 2 章

ネットワーク

2.1　グラフとネットワーク

ネットワークを解析するにあたり重要なのは，グラフ理論[7]である．グラフ理論は，数学の一分野である．ノード（節点・頂点）の集合とエッジ（枝・辺）の集合で構成されるグラフの性質について研究する学問である．グラフ理論における「グラフ」はノードとエッジの集合である．グラフ G はノードの集合 $V = \{v_1, v_2, \ldots, v_n\}$ とエッジの集合 $E = \{e_1, e_2, \ldots, e_n\}$ によって記述される．ノード i に繋がっているエッジの本数をそのノードの次数d_iという．ノードとエッジのパターンを変えることによって様々な特性を持つグラフが生成される．本研究では，ネットワークをグラフに置き換えて分析に用いる．

2.2　有向グラフと無向グラフ

・有向グラフ

向きのあるエッジとノードからなるネットワークのこと．World Wide Web（WWW）や食物連鎖を表すネットワークなどがこれにあたる．

例として，World Wide Web（WWW）を挙げる．WWW はウェブページをノード，ページ間のハイパーリンク関係をエッジとする巨大なネットワークである．WWW の中のあるウェブページ A から別のウェブページ B へのハイパーリンク関係は，「A が B を参照するのであって，B が A を参照するのではない」という意味で対象ではない．したがって，WWW 上でのネットサーフィンでは，

ページ A からページ B に行くことはできても，その逆はたどれない[7,8].

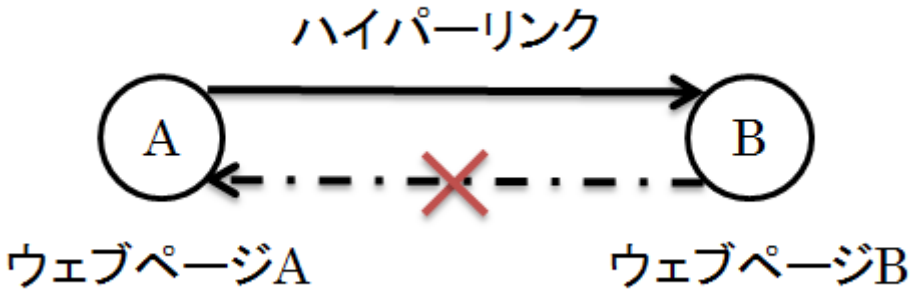

図 2.1　WWW の概念図

・無向グラフ

向きの属性を持たないエッジとノードからなるネットワークのこと．知人関係のネットワークなどがこれにあたる．

例えば，整数をノードとし，整数同士が互いに素である関係をエッジとする図 2.2[8]のようなネットワークにおいては, エッジが表す関係に向きをつけることはできない.

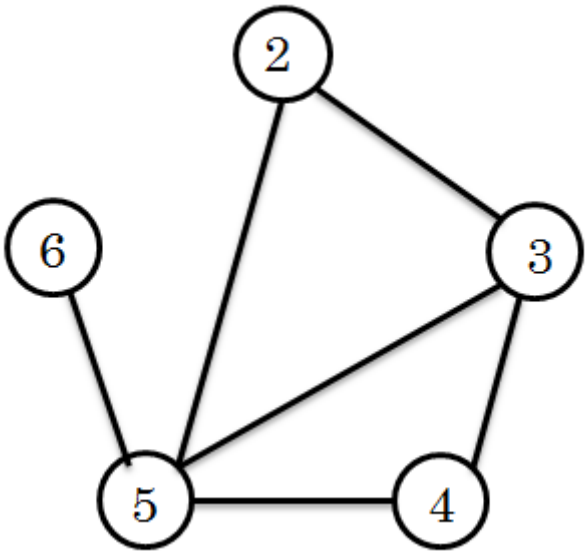

図 2.2　互いに素である整数のネットワーク

2.3　次数

2.1 節でも述べたが，ノードに繋がっているエッジの本数をそのノードの次数という．次数は，ネットワーク構造を議論する際に極めて重要となる基本的な量である．

2.3.1　入次数と出次数

ノードに入ってくるエッジの本数を**入次数**といい，出ていくエッジの本数を**出次数**という．

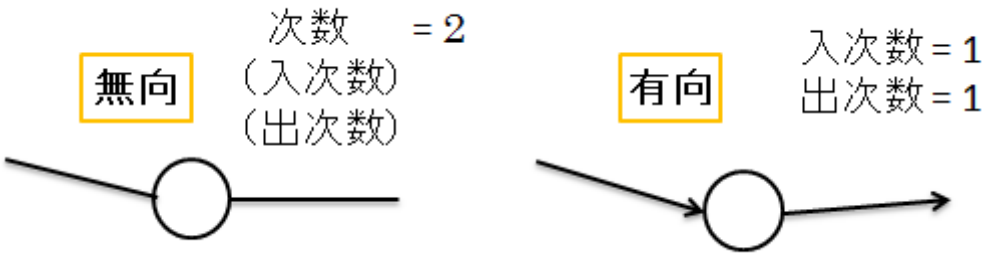

図 2.3　無向グラフ，有向グラフに関する次数

図 2.3 の有向グラフは入ってくるエッジが 1 本，出ていくエッジが 1 本であるので，入次数 1，出次数 1 となる．一方，無向グラフはエッジに向きがないため，双方向に移動が可能である．したがって，無向グラフに入次数，出次数という概念はないが，あえて入次数，出次数を表すならば，入次数 2，出次数 2 となる．

2.3.2　平均次数

次数を全ノードで平均した平均次数$\langle d \rangle$は，ネットワーク全体の大域的性質を表す量の一つとして重要である．$\langle d \rangle$を式で表せば，

$$\langle d \rangle = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{2.1}$$

と書ける．なお，Nは全ノード数である．
本研究では，シミュレーションに使うネットワークデータの次数平均の真値を式(2.1)で求める．

平均次数$\langle d \rangle$が全ノード数Nと同程度の大きさであるようなネットワークは，エッジ密度の非常に高い，密なネットワークであるといえる．逆に，$\langle d \rangle \ll N$なるネットワークは，エッジ密度の低い疎な繋がりのネットワーク構造を持つ．したがって，平均次数はネットワークのエッジの”密度”を反映した量であるといってもよい．

表 2.1[8]でいくつかの現実ネットワークのノード数と平均次数を挙げる．表 2.1 を見ると，平均次数$\langle d \rangle$はいずれもノード数Nに比べてはるかに小さいことがわかる．表 2.1 に挙げた例に限らず，身の周りのほとんどすべての複雑ネットワークの$\langle d \rangle$は 100 以下であり，$\langle d \rangle \ll N$が成立している．このように，実際のネットワークは，エッジ密度の低い疎なネットワークになっている．

表 2.1　現実の複雑ネットワークのノード数と平均次数

ネットワーク	ノード数 N	平均次数<d>
欧州鉄道網	4852	2.4
脳機能ネットワーク	31503	13.4
電子メールネットワーク	59812	2.88
俳優共演関係	449913	113.43
インターネット	228263	2.8
航空網	3880	9.7

また，次数の平均値だけでなく，平均の周りのばらつき具合（変動係数，分散，標準偏差等）や最大・最小次数など，次数の分布を知ることにより，ネットワークの平均構造を超えたより詳細な情報を得ることができる．

2.4　スケールフリー性

スケールフリー性[9]は，現実世界のネットワークが持つ性質である．一部のノードがほかのたくさんのノードとエッジでつながっており，大きな次数を持っている一方で，大多数のノードはごくわずかなノードとしかつながっておらず，次数は小さいという性質である．次数の大きなノードは「ハブ」とも呼

ばれる．

スケールフリー性は，社会学をはじめとするこれまでの研究から，現実世界のネットワークに幅広く観察されている．例えば，人々の持っている知人関係の数をみると，一部の人は非常にたくさんの知人を持っているが，大多数の人々の知人の数は限られている．WWW ではごく少数の有名サイトが数百万単位のリンクを集めているが，大多数のサイトはわずかなリンク先からしかリンクされていない．

現実に存在する様々なネットワーク（複雑ネットワーク）の多くは，スケールフリー性を持つことが知られている．このスケールフリー性を持つネットワークのことを，スケールフリーナネットワークという．

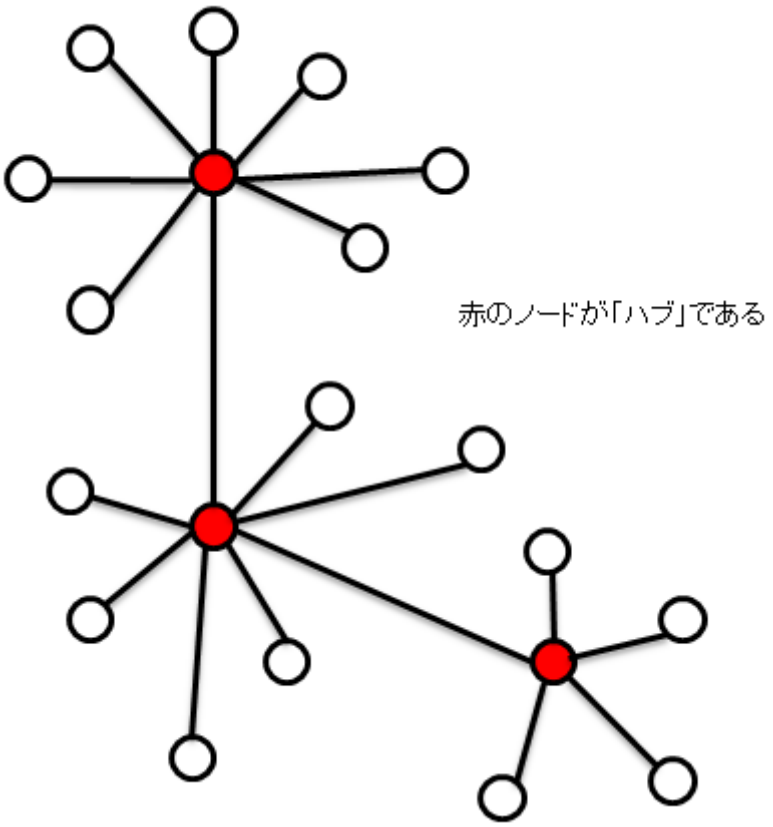

図 2.4　スケールフリーネットワークの例

第 3 章

ランダムウォークサンプリング

3.1　古典的ランダムウォークサンプリング

N個のノードを持つ無向連結グラフ上でのランダムウォークサンプリングを考える．古典的ランダムウォークサンプリングは隣接ノードの一つを等確率に選択し遷移しながらデータを収集する手法である．ノード i からノード j への推移確率を $p(i,j)$ とし，ノードiの次数をd_iとすると，$p(i,j)$は以下で与えられる．

$$p(i,j) = \begin{cases} \frac{1}{d_i}, & \text{ノード}j\text{が}i\text{に隣接する} \\ 0. & \text{その他} \end{cases} \tag{3.1}$$

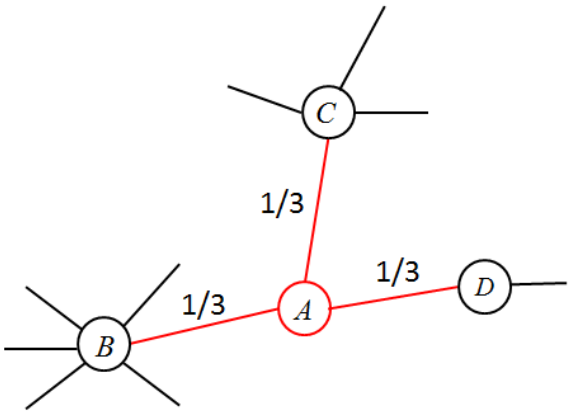

図 3.1　古典的ランダムウォークの遷移確率

図 3.1 は古典的ランダムウォークの遷移の例である．ノード A から隣接ノード B, C，D に遷移するとき，ノード A の次数は 3 であるため、ノード B，C，D への遷移確率は 1/3 となる．このようにすべて等確率で遷移する．

しかしながら，ネットワークのノードが異なる次数を持つときに，上述した手法をとると，多くのノードとリンクで繋がっている高次数ノードが遷移先に選ばれやすい傾向がある．一般的に，各ノードの訪問確率はノードの次数に比例することが知られている[1]．したがって古典的ランダムウォークサンプリングを用いた場合，高次数ノードのデータを偏って収集する結果となり，収集されたデータにバイアスが発生する．

3.2　Metropolis-Hastings アルゴリズム

Metropolis-Hastings アルゴリズム[2]は, ウォーカーが全てのノードを等確率で訪問するように，ウォーカーの隣接ノードへの遷移確率を，隣接ノードの次数に応じて調整する手法である．Metropolis-Hastings アルゴリズムではノード i からノード j への推移確率 $p(i,j)$ を以下の式で与える．

$$(i,j) = \begin{cases} \frac{1}{d_i}\min\left\{1, \frac{d_i}{d_j}\right\}, & \text{ノード}j\text{が}i\text{に隣接する} \\ 1 - \sum_{j \neq i} p(i,j), & j = i \\ 0. & \text{その他} \end{cases} \tag{3.2}$$

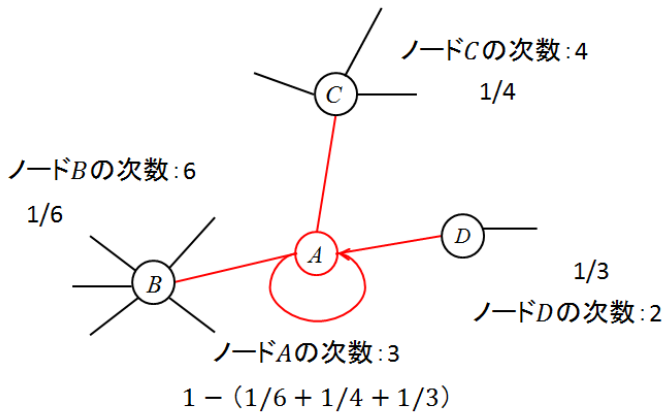

図 3.2　Metropolis-Hastings アルゴリズムの遷移確率

図 3.2 は Metropolis-Hastings アルゴリズムの遷移の例である．ノード A から隣接ノード B，C，D に遷移するとき，ノード A からノード B への遷移確率はノード B の次数が 6 より，1/6 となる．同様にノード A からノード C への遷移確率は 1/4 となる．ノード A からノード D への遷移確率はノード D の次数が 2 であるため，ノード A の次数 3 より小さい．したがって，1/3 となる．ノード A に留まる確率は 1−(1/6+1/4+1/3)となる．

Metropolis-Hastings アルゴリズムは，現在のノードをもう一度選択しなおすことで，現在のノードに留まったまま連続サンプリングを行うことを許容している．

3.3　特定のノード群を高頻度に訪問するランダムウォークサンプリング

ノード数Nを持つ無向連結グラフ上でのランダムウォークサンプリングを考える．π_iをウォーカーが定常状態でノードiに存在する確率とする．今，ウォーカーの定常分布$\{\pi_i\}_{i=1}^{N}$が以下の式に従うランダムウォークサンプリングを実現したいとする．

$$\pi_i = \rho(d_i)/N \tag{3.3}$$

ここで$\rho(x)$は任意の関数とする．Metropolis-Hastings アルゴリズムを拡張することにより，ノードiからノードjへの遷移確率$p(i,j)$を次の式で与えることで，訪問確率が式(3.3)に従うことがわかる．

$$p(i,j) = \begin{cases} \dfrac{1}{d_i}\min\left\{1, \dfrac{d_i\rho(d_j)}{d_j\rho(d_i)}\right\} & \text{ノード}j\text{が}i\text{に隣接する} \\ 1 - \sum_{j\neq i} p(i,j), & j = i \\ 0. & \text{その他} \end{cases} \tag{3.4}$$

特に，$\rho(d_i) = cd_i^n$（cは正規化定数）とすると，

$$p(i,j)=\begin{cases}\dfrac{1}{d_i}\min\left\{1,\dfrac{d_j^{n-1}}{d_i^{n-1}}\right\} & \text{ノード}j\text{が}i\text{に隣接する}\\ 1-\sum_{j\neq i}p(i,j), & j=i\\ 0. & \text{その他}\end{cases} \tag{3.5}$$

ここでnはランダムウォークのバイアスを調整するパラメータと考えることができる． $n>0$のとき，ウォーカーは高次数ノードを優先的に訪問し，$n<0$のとき，低次数ノードを優先的に訪問する．また，$n=1$のとき，古典的ランダムウォークとなり，$n=0$のとき，通常の Metropolis-Hastings アルゴリズムと一致する．

3.4　データサンプリング後のバイアス除去

古典的ランダムウォークサンプリングや 3.3 節で述べたサンプリング手法では，サンプリングで得られた情報にバイアスが含まれるので，事後にこのバイアスを除去する必要がある．このバイアス処理は確率測度の変換として定式化できる[4,5]．以下，3.3 節のランダムウォークサンプリングを例にとって説明する．二つの確率測度P_R，Pを考える．（$P[A]$はノードを均一にサンプリングした際に事象Aを見る確率，$P_R[A]$はランダムウォークサンプリングの際に事象Aを見る確率）．式(3.3)から，二つの確率測度の間には以下の関係が成立する．

$$P_R[A\cap\{D=k\}]=\rho(k)P[A\cap\{D=k\}] \tag{3.6}$$

ここで，Dは訪問したノードの次数を表す確率変数である．
式(3.6)のAを$A=\Omega$（全体集合）とすると，以下の式を得ることができる．

$$P_R[D=k]=\rho(k)P_R[D=k] \tag{3.7}$$

ここでの$\rho(k)$は次数に依存するバイアスに相当する．確率測度P（P_R）に対する期待値を$E[\cdot]$（$E_R[\cdot]$）とする．$\rho(k)=ck^n$のとき，

$$c^{-1}E_R[D^{-n}]=E_R\left[\frac{1}{\rho(D)}\right]=\sum\frac{1}{\rho(k)}P[D=k]=1$$

正規化定数cは$E_R[D^{-n}]$で与えられる. 式(3.6)から任意の確率変数Xと任意の関数$f(x)$について，次が成り立つ.

$$E[f(X)] = E_R[f(X)(\rho(D))^{-1}] \tag{3.8}$$

例えば、確率変数 X は次数，クラスタリング係数，Twitter の 1 週間での Tweet 数等に対応する．式(3.8)より次が成立する.

$$E[X] = E_R\left[\frac{X}{\rho(D)}\right] \qquad P[X = x] = E_R\left[\frac{\mathbf{1}(X = x)}{\rho(D)}\right] \tag{3.9}$$

ここで$\mathbf{1}(A)$は，Aが真であれば 1，偽であれば 0 をとる indicator 関数である．特に，$\rho(D) = E_R[D^{-n}]D^n$のとき,

$$E[X] = \frac{E_R[XD^{-n}]}{E_R[D^{-n}]}$$
$$P[X = x] = \frac{E_R[\mathbf{1}(X = x)D^{-n}]}{E_R[D^{-n}]} \tag{3.10}$$

となる.$n > 0$のとき, ランダムウォークは高次数ノードに優先的に訪問し,$n < 0$のとき，低次数ノードを優先的に訪問する．また，$n = 1$のとき，古典的ランダムウォークとなり，$n = 0$のとき，通常の Metropolis-Hastings アルゴリズムと一致する． 式(3.10)を用いれば，3.3 節のランダムウォークサンプリングで得られたXに関する情報から，Xの真の期待値や分布を求めることができる.

第 4 章

複数のウォーカーによるランダムウォークサンプリング

4.1　Multiple ランダムウォーク

Multiple ランダムウォークサンプリングは，初期状態で複数のウォーカーをランダムなノードに配置し，以降，それぞれを独立に動かしながら，情報を収集する手法である．

ステップ数が大きく取れない場合，ランダムウォークサンプリングで収集した情報にはウォーカーの初期位置に依存するバイアスが残るが，複数のウォーカーをランダムに抽出したノードに初期配置することで，ウォーカーの初期位置に依存するバイアスを緩和することができる．ただし，3.4 節で述べたバイアス除去法はウォーカーが特定のノードに偏って存在することを仮定しているので，ウォーカーをランダムに初期配置して情報を収集し，3.4 節のバイアス除去法を適用すると，それがかえってバイアスを生む結果となる．

図 4.1 はウォーカーが 3 つのときの Multiple ランダムウォーク（パラメータ $n = 0$のとき）の例である．遷移確率はウォーカーごとに 3.3 節の式(3.5)に従う．

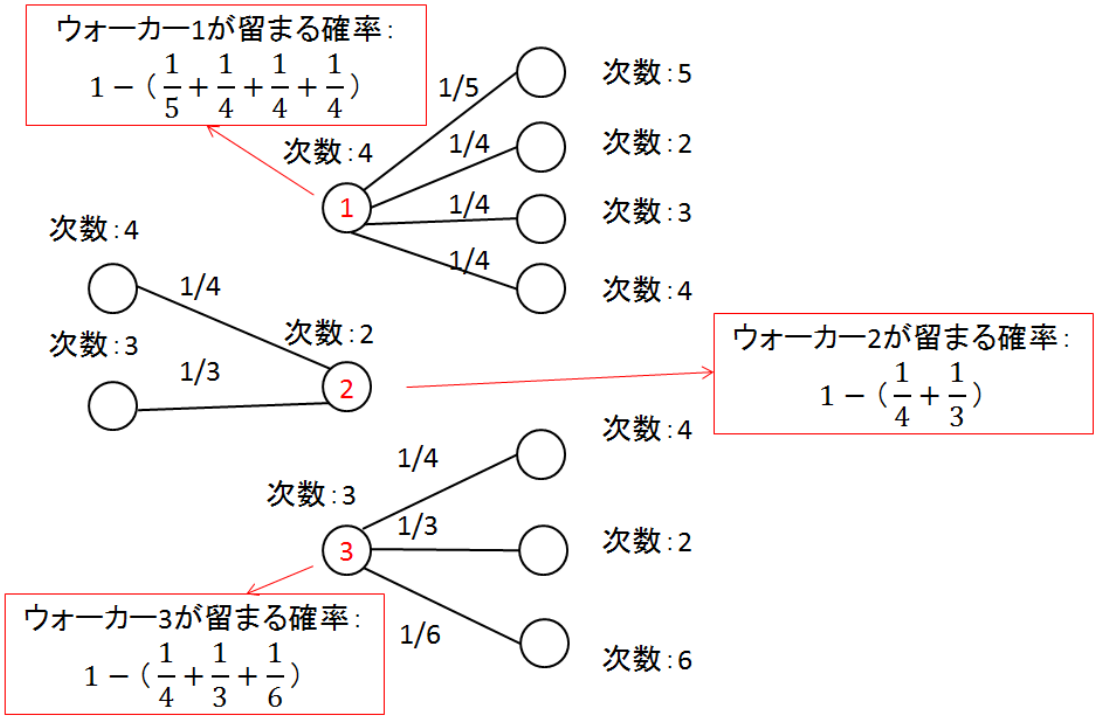

図 4.1　Multiple ランダムウォークの遷移確率

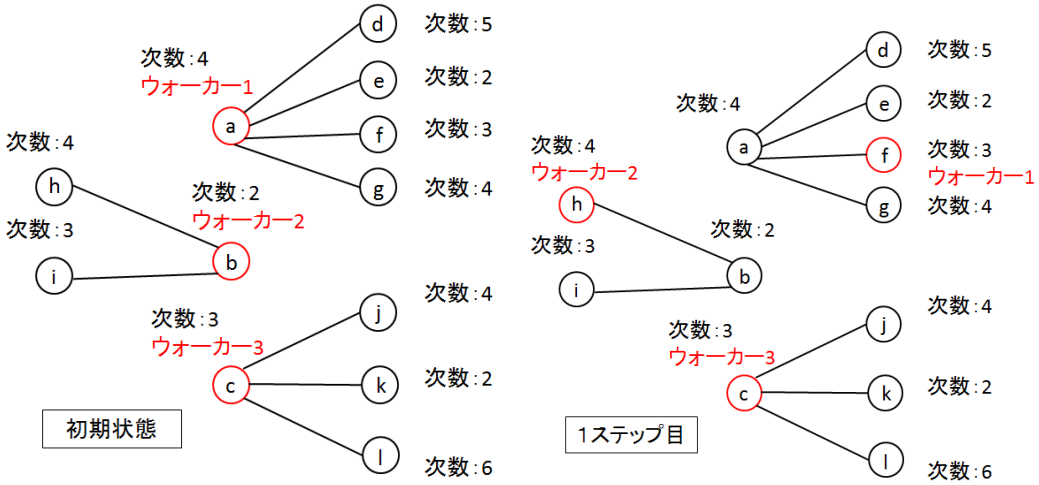

図 4.2　Multiple ランダムウォークの遷移

図 4.2 は Frontier Sampling の遷移の例である．
初期状態でウォーカー1，2，3 はそれぞれ a，b，c のノードに存在する．
1 ステップ目でウォーカー1 がノード f に遷移, ウォーカー2 がノード h に遷移, ウォーカー3 が留まる．
以上の条件のとき，3.4 節のバイアス除去法を適用し，求める確率変数 X に次数

をとると、平均次数を求めることができる．（$X = D$となる．）

$$\text{平均次数} = \frac{(4+2+3)^{-n+1} + (3+4+3)^{-n+1}}{(4+2+3)^{-n} + (3+4+3)^{-n}}$$

となる．

4.2 Frontier Sampling

Frontier Sampling[3]は，複数のウォーカーの中から一つのウォーカーを選んで，その一つを隣接ノードに遷移させるアルゴリズムである．ウォーカーの数をmとし，i番目のウォーカーが存在するノードの番号$w(i)$とする．Frontier Sampling において，i番目のウォーカーが遷移対象として選ばれる確率は

$$p_{walker}(i) = \frac{d_{w(i)}}{d_{w(1)} + \cdots + d_{w(m)}} \tag{4.1}$$

選択されたウォーカーは隣接ノードのいずれかに等確率で遷移する．Frontier Sampling でのウォーカーの遷移は，m次元空間のマルコフ連鎖としてモデル化できる．定常状態においてウォーカーが$(w(1), w(2), \ldots, w(m))$に存在する確率は，ウォーカーが存在するノードの次数の和$d_{w(1)} + \cdots + d_{w(m)}$に比例する．Frontier Sampling は複数のウォーカーを独立に動かす場合よりも速やかに定常状態に収束し，ウォーカーの初期位置に依存するバイアスを緩和できることが証明されている[3].

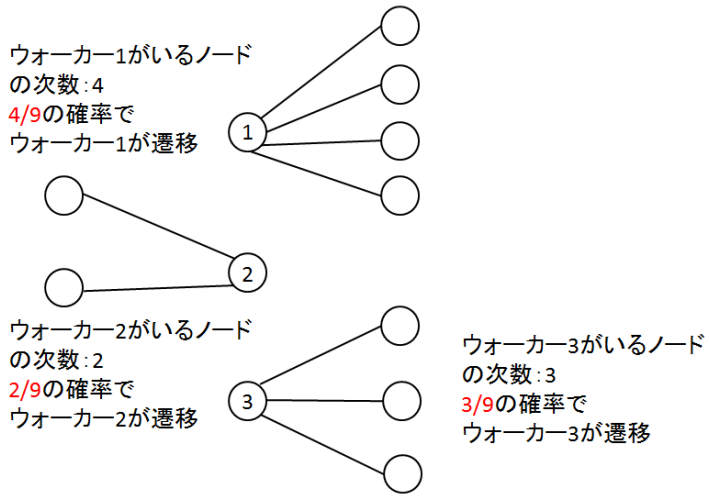

図 4.3　Frontier Sampling のウォーカー選択確率

図 4.3 はウォーカーが 3 つのときの Frontier Sampling の例である．ウォーカー1，2，3 が存在するノードの次数はそれぞれ 4，2，3 である．したがって，ウォーカー1，2，3 の選択確率はそれぞれ 4/9，2/9，3/9 となる．その後，選択されたウォーカーは隣接ノードのいずれかに等確率で遷移する．

本研究では Frontier Sampling に 3.3 節で説明したウォーカー制御法を組み合わせて，複数のウォーカーを意図的に特定のノード群を高頻度に訪問させて，情報を収集させることを可能とするサンプリング手法を提案する．提案手法において，ウォーカー群が$\vec{w} = (w(1), w(2), \ldots, w(m))$から，

$\overrightarrow{w'} = (w'(1), w'(2), \ldots, w'(m))$に遷移する確率$p(\vec{w}, \vec{w}')$は以下で与えられる．

$$p(\vec{w}, \vec{w}') = \begin{cases} \dfrac{1}{d_{\vec{w}}} \min\left\{1, \dfrac{d_{\overrightarrow{w'}}^{n-1}}{d_{\vec{w}}^{n-1}}\right\} & \vec{w}が\overrightarrow{w'}に隣接する \\ 1 - \displaystyle\sum_{j \neq i} p(i,j), & \overrightarrow{w'} = \vec{w} \\ 0. & その他 \end{cases} \tag{4.2}$$

ここで，$d_{\vec{w}} = d_{w(1)} + \cdots + d_{w(m)}$，$d_{\overrightarrow{w'}} = d_{w'(1)} + \cdots + d_{w'(m)}$である．データ収集後，情報に含まれるバイアスを除去するために，3.4 節の手法を適用する．

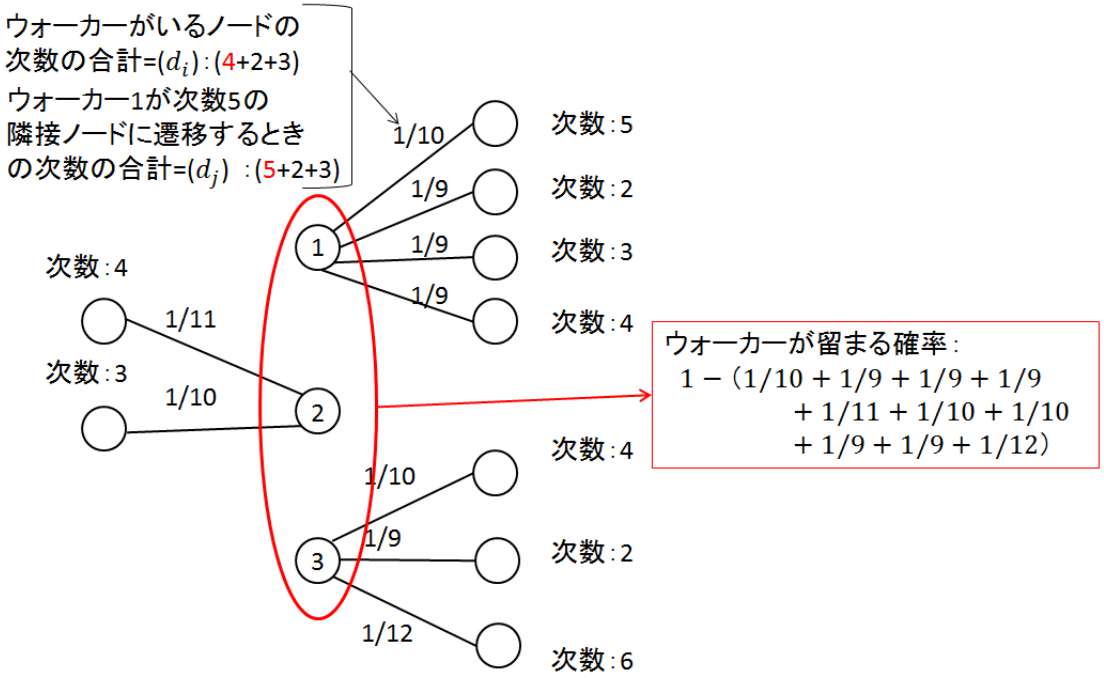

図 4.4　Frontier Sampling 拡張時のウォーカーの遷移確率

図 4.4 は式(4.2)の遷移確率に従う Frontier Sampling（パラメータ$n = 0$のとき）の例である．ウォーカー1 から次数 5 の隣接ノードに遷移する確率は 1/{5+2（ウォーカー2 の次数）+（ウォーカー3 の次数）} =1/10 となる．また，ウォーカー1 が次数 2 の隣接ノードに遷移する確率は，ウォーカー1 が現在いるノードの次数が遷移先のノードの次数より大きいため，1/9 となる．また、ウォーカーが現在のノードに留まる確率は {1- (すべてのウォーカーの隣接ノードへの遷移確率)} となる．

図 4.5 は Frontier Sampling の遷移の例である．

初期状態でウォーカー1，2，3 はそれぞれ a，b，c のノードに存在する．
1 ステップ目でウォーカー1 がノード f に遷移．
2 ステップ目でウォーカー3 がノード l に遷移．
3 ステップ目でウォーカーが現在いるノードに留まる．

以上の条件のとき，3.4 節のバイアス除去法を適用し，求める確率変数 Xに次数をとると、平均次数を求めることができる．（$X = D$となる．）

平均次数

$$= \frac{(4+2+3)^{-n+1} + (3+2+3)^{-n+1} + (3+2+6)^{-n+1} + (3+2+6)^{-n+1}}{(4+2+3)^{-n} + (3+2+3)^{-n} + (3+2+6)^{-n} + (3+2+6)^{-n}}$$

となる.

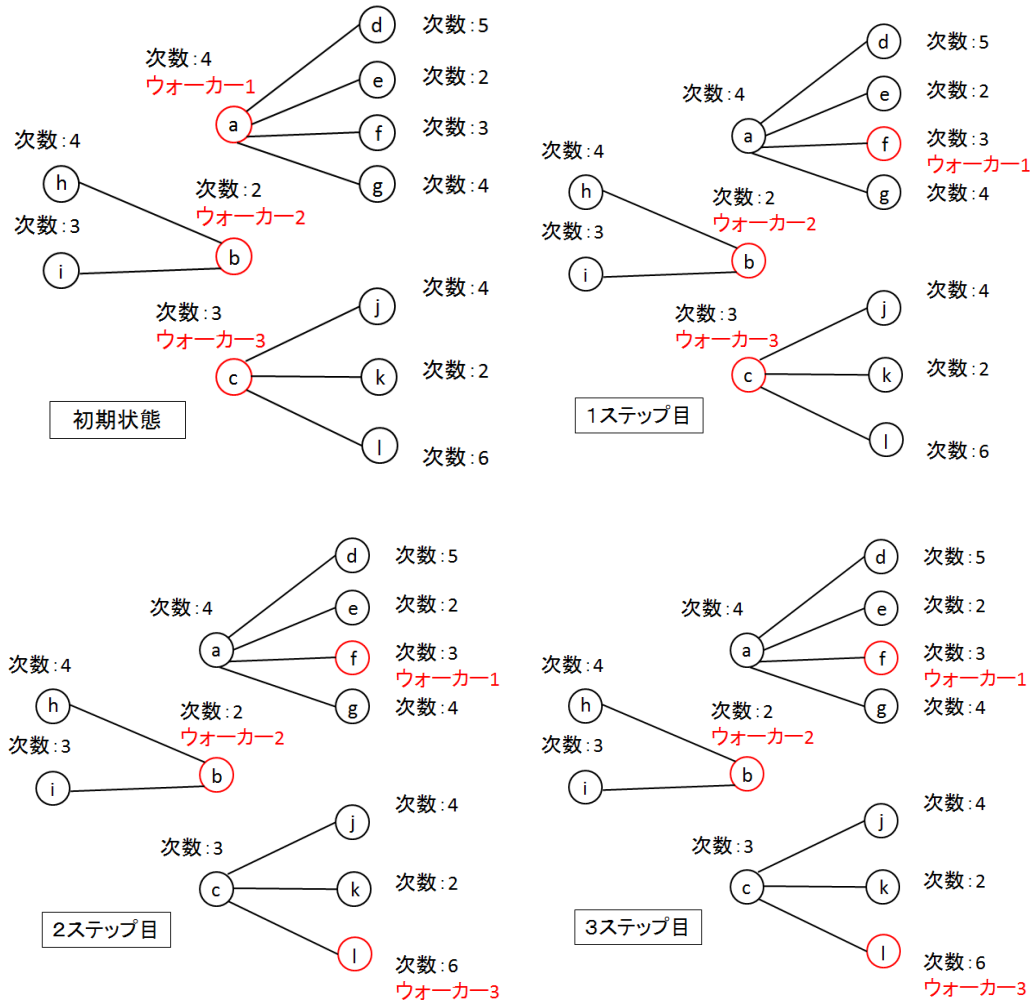

図 4.5　Frontier Sampling の遷移

第 5 章

シミュレーション実験

5.1　シミュレーション条件

表 5.1 の特徴を持つネットワークとインターネット上で公開されている 2 つのネットワーク（Facebook, Gnutella：表 5.2）のトポロジーデータ[6]を用いて，コンピュータ上に仮想的にネットワークを構築し，ネットワーク上でランダムウォークさせながら，訪問先のノードの次数データを収集させ，収集データに基づいて平均次数や次数分布を推定させた．平均次数や次数分布の推定には 3.4 節のバイアス除去法を用いた．なお，表 5.1 のネットワークは次数分布がおよそべき乗則に従うスケールフリーネットワーク（べき指数=2）である．

ランダムウォークサンプリングは
(1) Frontier Sampling
(2) Multiple ランダムウォークサンプリング
(3) Single ランダムウォークサンプリング
（ウォーカー数が 1 つのときのランダムウォークサンプリング）
の 3 通りを用いた．

これらのネットワークからデータを収集する際，Frontier Sampling の遷移確率である式(4.2)もしくは，Single (or Multiple)ランダムウォークの遷移確率である式(3.5)に含まれるバイアス調整パラメータn は-1, 0, 1, 2 に設定する．

シミュレーションでは，ウォーカーが一つのノードを訪問するたびに 1 に等しい Budget が必要であること，利用できる Budget の総量には制限あることを

仮定しておこなった．Budget の総量を B，ウォーカーの数を m とすると，Frontier Sampling では各ウォーカーの平均ステップ数は B/m に等しく（ウォーカーによって違いがある），Multiple ランダムウォークサンプリングでは各ウォーカーのステップ数は全て B/m に等しい．

表 5.1　ネットワークデータ

グラフ	ノード数	リンク数
無向	3000	10006
無向	10000	38367

表 5.2　ネットワークデータ

ネットワーク	グラフ	ノード数	リンク数
Facebook	無向	4039	88234
Gnutella	有向	6299	20776

なお，表 5.2 に用いたネットワークデータのうち，Gnutella はリンクに向きがある有向グラフであり，そのままではランダムウォークサンプリングを適用できない． そのため，有向グラフについてはシミュレーションを実施する前に無向化を行った． 具体的には，ノードiからノードj，もしくはjからiの向きに有向リンクが存在する場合は，$i-j$間に無向リンクが一本設定されているとした．（ノードiからノードの向きに有向リンクが存在しかつjからiの向きにも有向リンクが存在する場合も，$i-j$間に無向リンクが一本設定されているとする．）

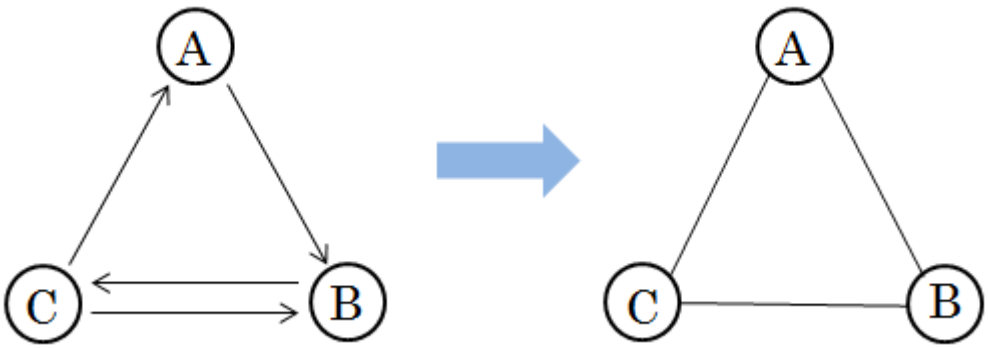

図 5.1　有向リンクの無向化

5.2　シミュレーション結果：Facebook

5.2.1　平均次数の推定

ランダムウォークサンプリングでは，ランダムウォークを開始する初期ノードによりサンプリングデータに偏りが発生する．そこで，Frontier Sampling，ランダムウォークサンプリング，Single ランダムウォークサンプリングの 3 通りのランダムウォークについて初期ノードを 100 通り変えて，平均次数の推定値を 100 通り求め，その平均を算出した．Budget をノード数の 2%からノード数に等しい値まで変えて，推定平均次数をプロットした結果を Facebook のパラメータごとに，図 5.2（$n = -1$：低次数ノード優先），図 5.3（$n = 0$：Metropolis-Hastings アルゴリズム），図 5.4（$n = 1$：古典的ランダムウォーク），図 5.5（$n = 2$：高次数ノード優先）に示す．横軸は，（ステップ数ではなく，）ステップ数のノード数に対する比，縦軸は推定された平均次数を表している．

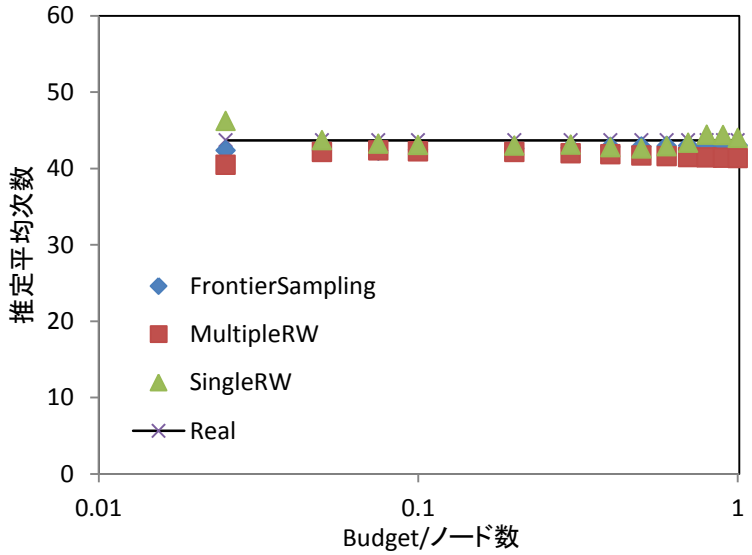

図 5.2　推定平均次数（$n = -1$：低次数ノード優先）

図 5.2 から，低次数ノード優先では，すべてのサンプリングで平均次数の真値と近い値を推定しているように読み取れる．しかし，Multiple ランダムウォークは，ステップ数が多くなるとともに真値との誤差も大きくなってしまう．

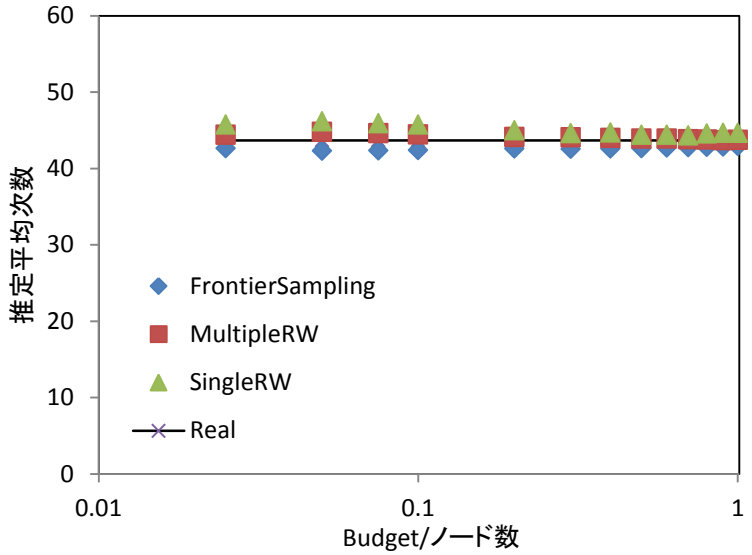

図 5.3　推定平均次数（$n = 0$：MH アルゴリズム）

図 5.3 から，Metropolis-Hastings アルゴリズムでも，すべてのサンプリングで平均次数の真値と近い値を推定しているように読み取れる．

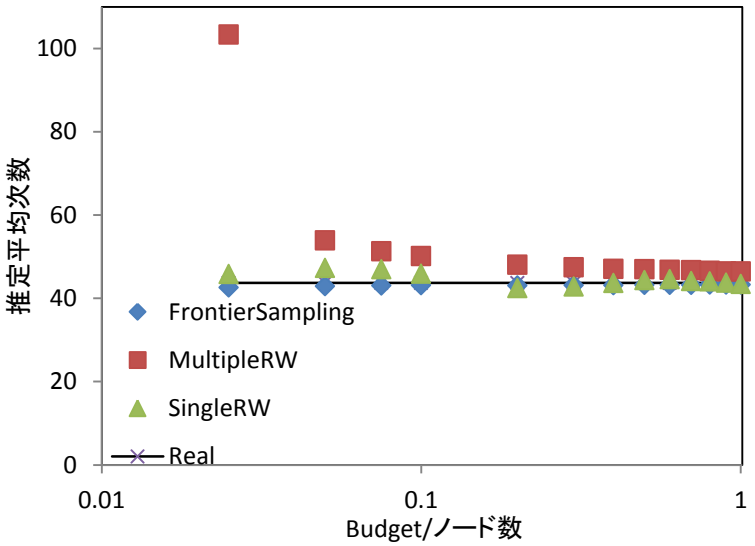

図 5.4　推定平均次数（$n = 1$：古典的ランダムウォーク）

図 5.4 から，古典的ランダムウォーク遷移では，　Frontier Sampling と Single ランダムウォークサンプリングは正確に平均次数を推定できているが，

Multiple ランダムウォークサンプリングは特に Budget が少なく，十分なステップ数が得られない場合，推定誤差が大きいことがわかる．

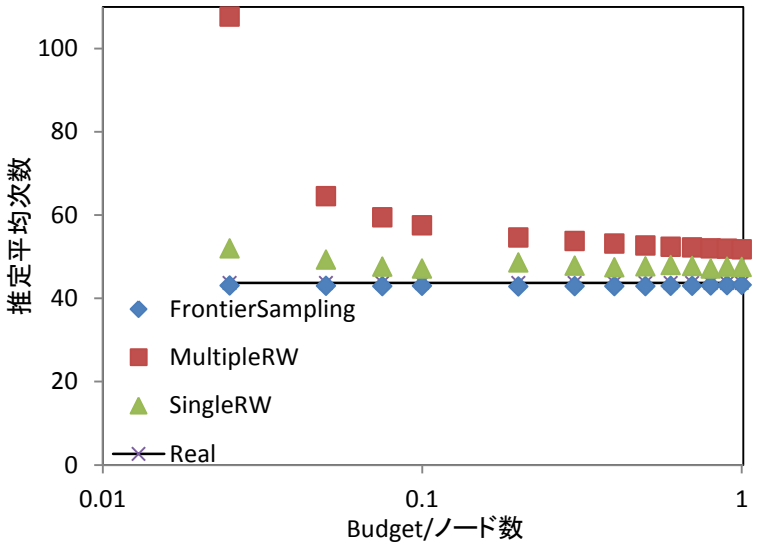

図 5.5　推定平均次数推定（$n = 2$：高次数ノード優先）

図 5.5 から，高次数ノード優先は，Multiple ランダムウォーク，Single ランダムウォークは真値と比べ誤差が生じていることが読み取れる．一方，Frontier Sampling では真値と近い平均次数を推定していることが読み取れる．

以上から，Multiple ランダムウォークサンプリングの推定誤差は，ウォーカーの初期配置のバイアスに起因していると考えられる．Frontier Sampling では，ウォーカーの初期配置のバイアスが速やかに減少する．Single ランダムウォークサンプリングでは（一つのウォーカーのステップ数を十分大きく取れるので）やはり初期配置バイアスの影響は小さい．したがって，Frontier Sampling がどんな遷移確率でも（全部のパラメータで）平均次数の真値を推定できるサンプリング方法であるといえる．

次に，少ない遷移回数で平均次数を推定できているか調べるために，ランダムウォークのステップ数がノード数の 10 分の 1 に達したときの平均次数の推定値を表 5.3，図 5.6 に示す．

表 5.3 平均次数

n	-1	0	1	2	真値
Frontier Sampling	42.46	42.38	43.14	42.88	43.69
Multiple RW	42.23	44.48	50.14	57.53	43.69
Single RW	43.07	45.73	45.92	47.04	43.69

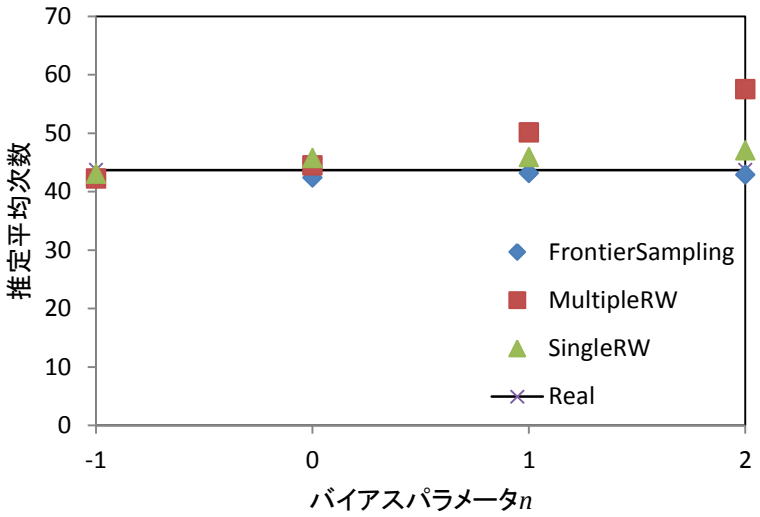

図 5.6 推定平均次数

表 5.3，図 5.6 より，Multiple ランダムウォークの平均次数の推定値は真値と比べて大きく誤差が出ている．すなわち、推定の精度が悪いことが分かる．また，パラメータの値が増えるほど、平均次数の推定値と真値の誤差も大きくなっている．原因として，ウォーカー各々の遷移回数が少ないことが考えられる．Single ランダムウォークとの比較から，遷移の回数が推定値の精度に関わってくることが分かる．

一方で，Frontier Sampling はすべてのネットワークかつすべてのパラメータでほぼ真値に近い値が推定できている．Single ランダムウォークも真値に近い推定値が存在するが、低次数優先，高次数優先ランダムウォークのときの推定値に誤差が見える．以上から，Frontier Sampling は少ない遷移回数で平均次数を推定できる効率的なサンプリング方法であることが分かる．

5.2.2　平均次数の変動係数

ランダムウォークサンプリングでは，ランダムウォークを開始する初期ノードによりサンプリングデータに偏りが発生する．そこで，初期ノードを 100 通り変えて，平均次数の推定値を 100 通り求め，ステップ数ごとの変動係数の値を算出し，初期ノードによるデータのばらつきを調べることで，Frontier Sampling，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングの 3 通りのランダムウォークのサンプリング精度の安定性を比較した．異なる 100 通りの初期ノードで Frontier Sampling，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングより推定した 100 通りの平均次数の変動係数の値を比較する．変動係数とは相対的なばらつきを表す指標で，標準偏差を$\sqrt{\sigma^2}$，算術平均を$\bar{x}$とおくと，変動係数$C.V$は以下で得られる．

$$C.V = \frac{\sqrt{\sigma^2}}{\bar{x}} \tag{5.1}$$

パラメータごとの比較結果を図 5.7（$n = -1$：低次数ノード優先），図 5.8（$n = 0$：MH アルゴリズム），図 5.9（$n = 1$：古典的ランダムウォーク），図 5.10（$n = 2$：高次数ノード優先）に示す．横軸は，（ステップ数ではなく，）ステップ数のノード数に対する比，縦軸は変動係数の値を表している．

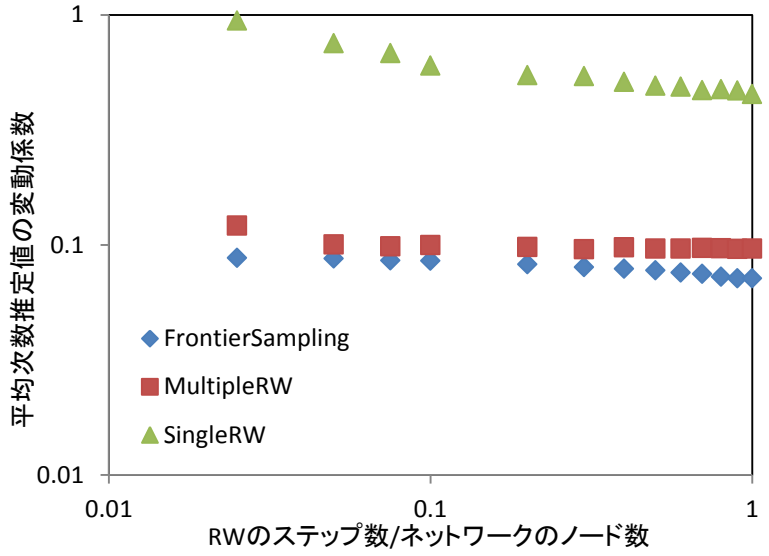

図 5.7　平均次数の変動係数（$n = -1$：低次数ノード優先）

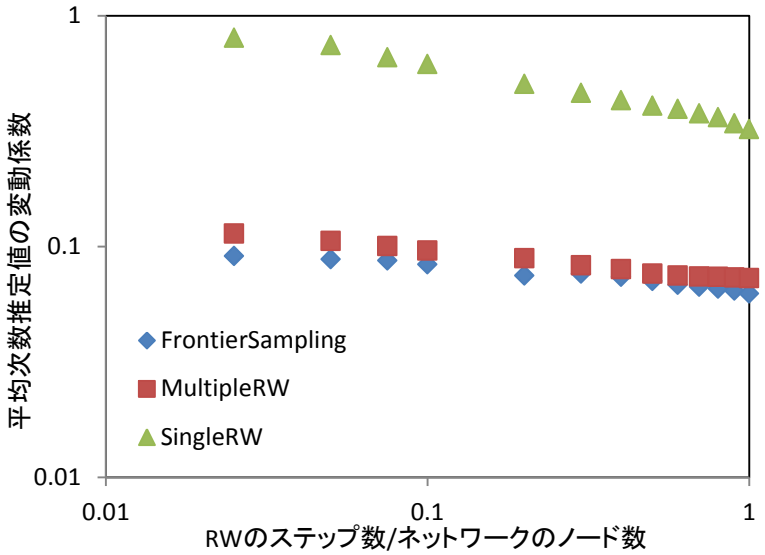

図 5.8　平均次数の変動係数（$n = 0$：MH アルゴリズム）

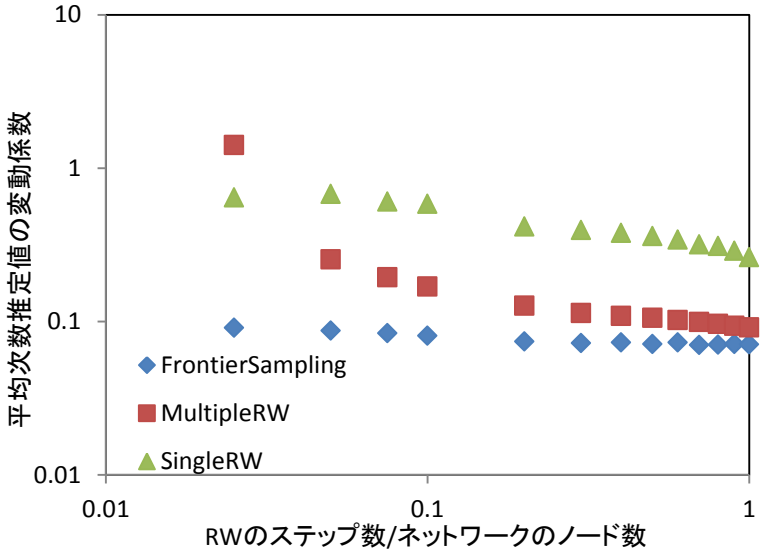

図 5.9　平均次数の変動係数（$n = 1$：古典的ランダムウォーク）

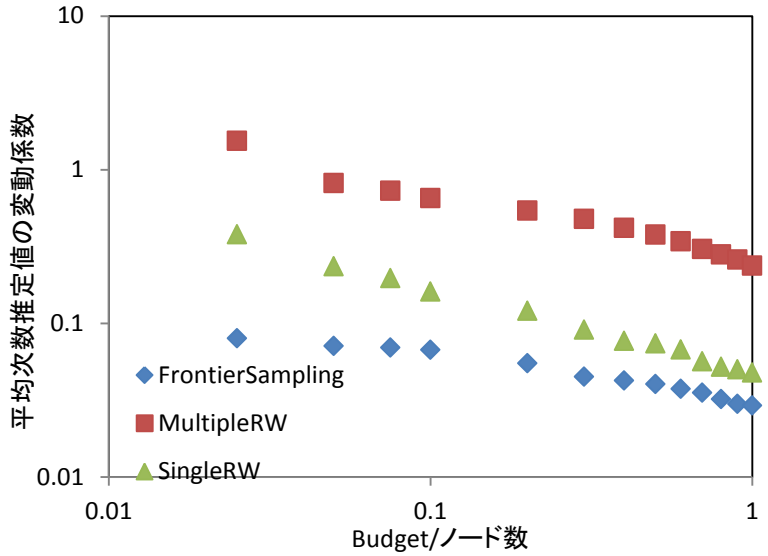

図 5.10　平均次数の変動係数（$n = 2$：高次数ノード優先）

図 5.7～図 5.10 に示されているように，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングと比較すると，Frontier Sampling はすべてのパラメータで変動係数の値が最も小さいことが分かる．すなわち，Frontier Sampling は初期ノードによるサンプリングデータのばらつきが小さく，安定した推定値を得ることができると言える．

5.2.3　次数分布の推定

Frontier Sampling，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングの 3 通りのランダムウォークについて初期ノードを 100 通り変えて，低次数ノード比率（次数が 5 以下のノードの比率），および高次数ノード比率（次数が上位 10%値を超えるノードの比率）を推定した結果を示す．高次数ノード比率を調べる際には，予め次数の上位 10%値をオフラインで求めておき，ウォーカーが訪問したノードの次数が上位 10%値を超えている頻度から高次数ノード比率を推定させることとした．一般に，低（高）次数ノード比率を推定する際には，低（高）次数ノードを優先訪問させることが望ましいと考えられるため，バイアス調整パラメータnは-1 から 2 まで変えて，結果を取得した．サンプリング Budget の総量はノード数の 1 割とした．

低次数ノード比率の推定結果を図 5.11 に示す．Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングはバイアスパラメータ依存性がある．とくに，Multiple ランダムウォークサンプリングのバイアスパラメータ依存性は顕著である．Multiple ランダムウォークサンプリングはパラメータが 0 を超えて大きくなると推定精度が劣化する．Single ランダムウォークサンプリングはパラメータが 1 を超えて大きくなった場合とパラメータが-1 の場合に推定精度が劣化した．一方，やや意外なことに，Frontier Sampling はバイアスパラメータ依存性がほとんど見られず，推定結果が Single，Multiple ランダムウォークサンプリングのいずれよりも安定していた．

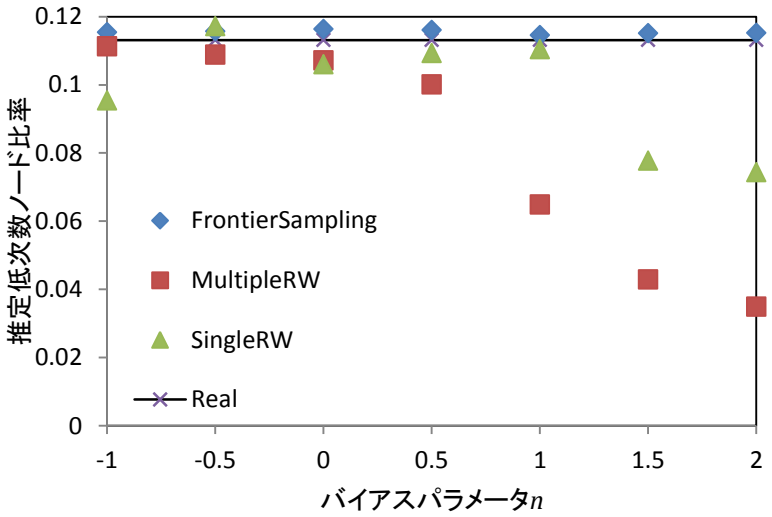

図 5.11　低次数ノード比率推定結果

次に，高次数ノード比率推定結果をプロットした結果を図 5.12 に示す．やはり Frontier Sampling はバイアスパラメータ依存性がほとんど見られず，推定結果が Single，Multiple ランダムウォークサンプリングのいずれよりも安定していた．また，意外なことに，Multiple ランダムウォークサンプリングについては，高次数ノード比率を推定する際にも，バイアスパラメータが低いほど（低次数優先訪問とするほど）推定結果が向上した．

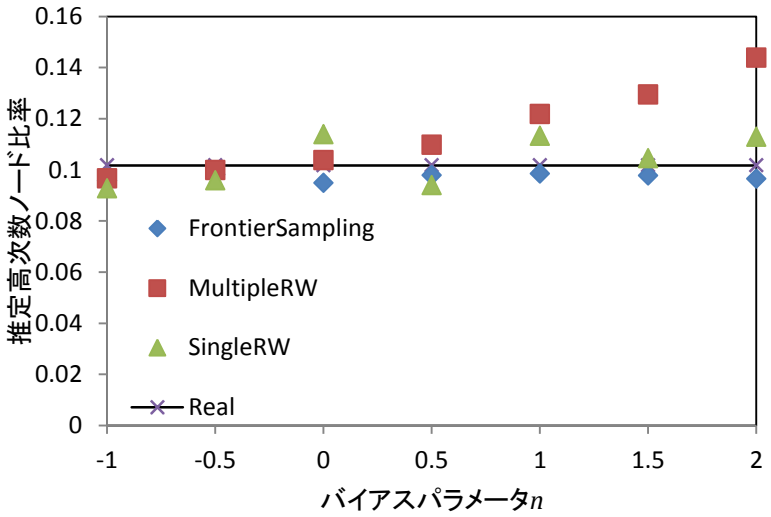

図 5.12　高次数ノード比率推定結果

5.3　シミュレーション結果：Gnutella

5.3.1　平均次数の推定

Frontier Sampling，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングの 3 通りのランダムウォークについて初期ノードを 100 通り変えて，平均次数の推定値を 100 通り求め，その平均を算出した．Budget をノード数の 2%からノード数に等しい値まで変えて，推定平均次数をプロットした結果を Gnutella のパラメータごとに，図 5.13（$n = -1$：低次数ノード優先），図 5.14（$n = 0$：Metropolis-Hastings アルゴリズム），図 5.15（$n = 1$：古典的ランダムウォーク），図 5.16（$n = 2$：高次数ノード優先）に示す．横軸は，（ステップ数ではなく，）ステップ数のノード数に対する比，縦軸は推定された平均次数を表している．

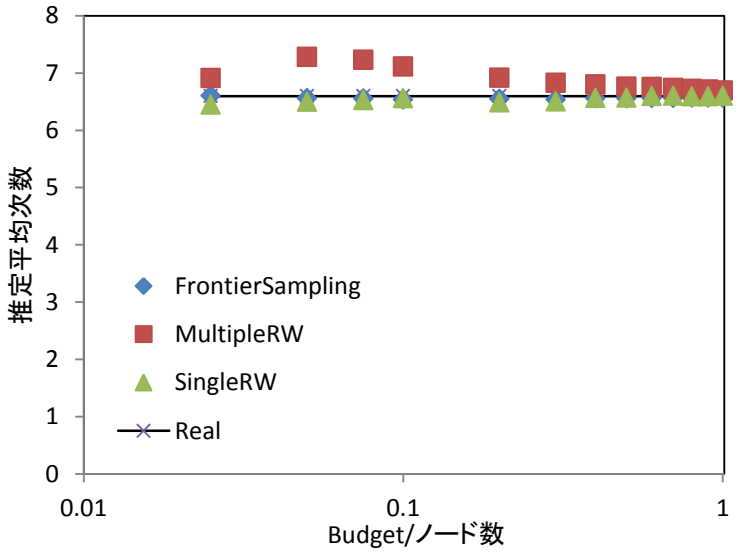

図 5.13　推定平均次数（$n = -1$：低次数ノード優先）

図 5.13 から，低次数ノード優先遷移確率では，すべてのサンプリングで平均次数の真値と近い値を推定しているように読み取れる．しかし，Multiple ランダムウォークは，ステップ数が小さいときに真値との誤差が見える．

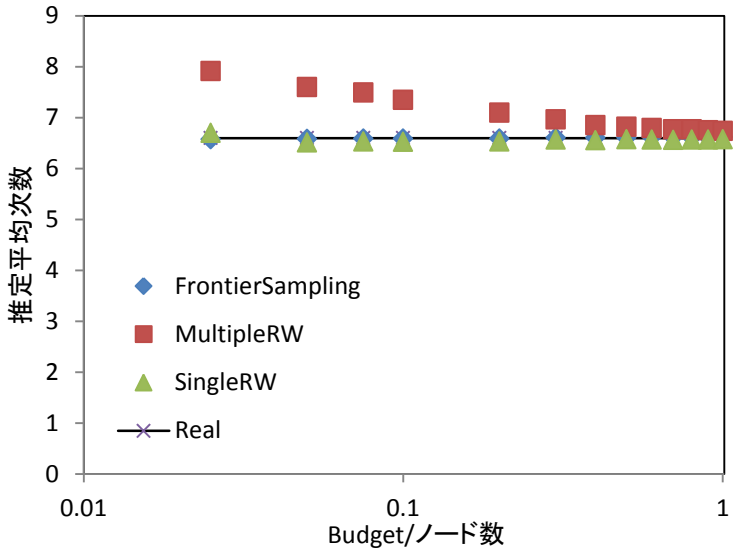

図 5.14　推定平均次数（$n = 0$：MH アルゴリズム）

図 5.14 から，Metropolis-Hastings アルゴリズムでも，すべてのサンプリングで平均次数の真値と近い値を推定しているように読み取れる．しかし，低次数ノード優先遷移確率と同様に，Multiple ランダムウォークは，ステップ数が小さいときに真値との誤差が見える．また，低次数ノード優先遷移のときよりも誤差が大きくなっている．

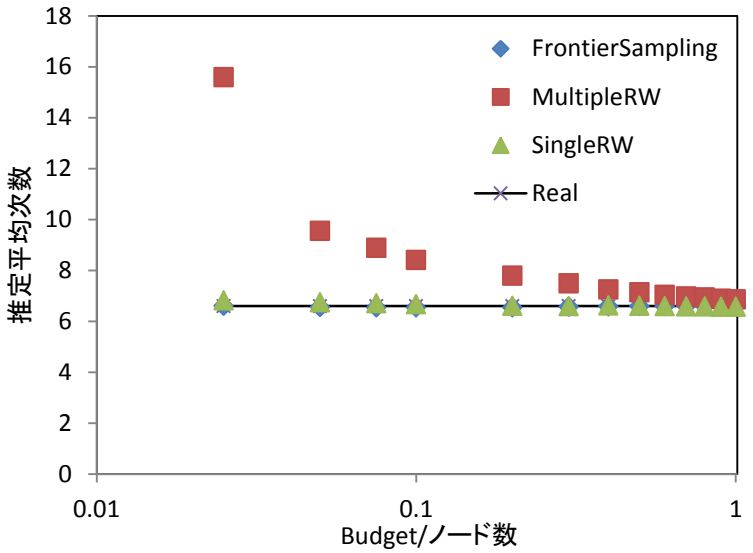

図 5.15　推定平均次数（$n = 1$：古典的ランダムウォーク）

図 5.15 から，古典的ランダムウォークでも同様に，Multiple ランダムウォークは，ステップ数が小さいときに真値との誤差が見える．また，さらに誤差が大きくなっている．Frontier Sampling と Single ランダムウォークでは，真値と近い値を推定していることが読み取れる．

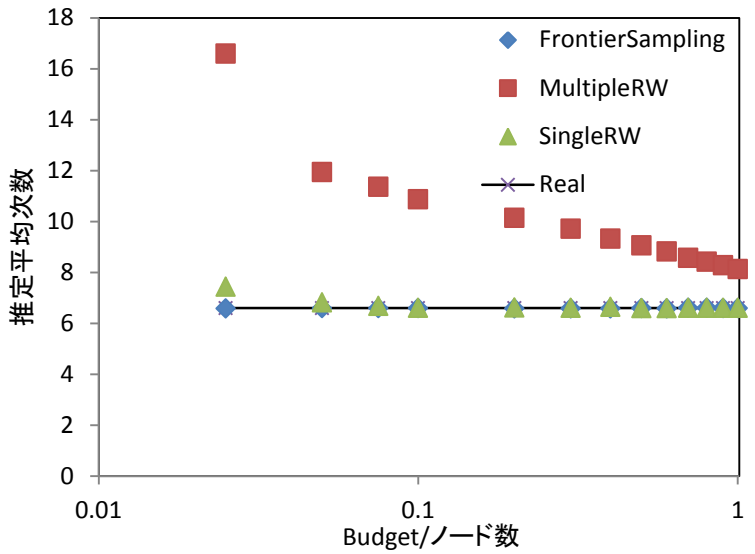

図 5.16　推定平均次数（$n = 2$：高次数ノード優先）

図 5.16 から，高次数ノード優先も同様に，Multiple ランダムウォークは，ステップ数が小さいときに真値との誤差が見える．また，さらに誤差が大きくなっている．一方，Frontier Sampling ，Single ランダムウォークでは真値と近い平均次数を推定していることが読み取れる．

Facebook のときと比べて，Gnutella では Single ランダムウォークサンプリングの推定の精度が良く見える．これは，Facebook に比べ，Gnutella はリンク数が少なく，平均次数も小さいため，低次数，高次数ノードに優先的に遷移させても影響が少なかったためだと考える．

以上から，Frontier Sampling がどんな遷移確率でも（全部のパラメータで）平均次数の真値を推定できるサンプリング方法であるといえる．

次に，少ない遷移回数で平均次数を推定できているか調べるために，ランダムウォークのステップ数がノード数の 10 分の 1 に達したときの平均次数の推定値を表 5.4，図 5.17 に示す.

表 5.4 平均次数

n	-1	0	1	2	真値
Frontier Sampling	6.542	6.58	6.543	6.581	6.597
Multiple RW	7.11	7.348	8.4	10.88	6.597
Single RW	6.55	6.52	6.66	6.61	6.597

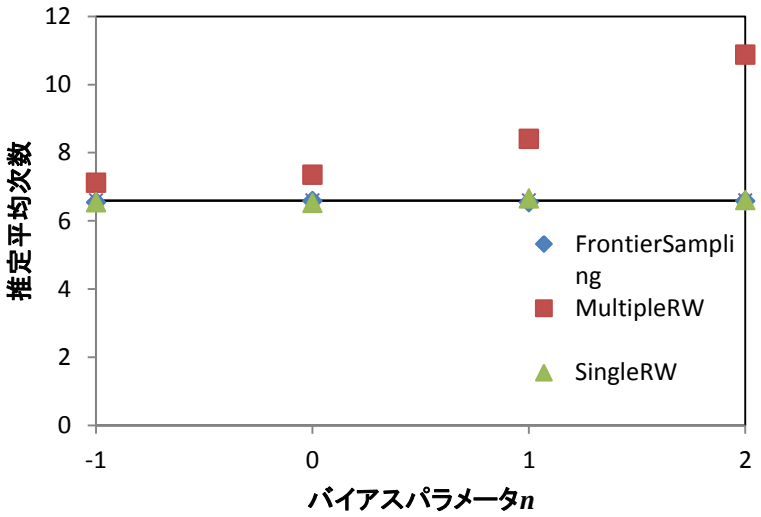

図 5.17 推定平均次数

表 5.4，図 5.17 より，Multiple ランダムウォークの平均次数の推定値は真値と比べて大きく誤差が出ている．すなわち、推定の精度が悪いことが分かる．また，パラメータの値が増えるほど，平均次数の推定値と真値の誤差も大きくなっている．原因として，ウォーカー各々の遷移回数が少ないことが考えられる．Single ランダムウォークとの比較から，遷移の回数が推定値の精度に関わってくることが分かる．

一方で，Frontier Sampling，Single ランダムウォークはすべてのネットワークかつすべてのパラメータでほぼ真値に近い値が推定できている．Single ランダムウォークの精度が良い理由は上述した通り，リンク数が少なく，平均次数

も小さいネットワークであるためだと考える．以上から，Frontier Sampling は少ない遷移回数で平均次数を推定できる効率的なサンプリング方法であることが分かる．

5.3.2　平均次数の変動係数

初期ノードを 100 通り変えて 平均次数の推定値を 100 通り求め，ステップ数ごとの変動係数の値を算出し，初期ノードによるデータのばらつきを調べることで，Frontier Sampling，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングの 3 通りのランダムウォークのサンプリング精度の安定性を比較した．異なる 100 通りの初期ノードで Frontier Sampling，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングより推定した 100 通りの平均次数の変動係数の値を比較する．パラメータごとの比較結果を図 5.18（$n = -1$：低次数ノード優先），図 5.19（$n = 0$：MH アルゴリズム），図 5.20（$n = 1$：古典的ランダムウォーク），図 5.21（$n = 2$：高次数ノード優先）に示す．横軸は，（ステップ数ではなく，）ステップ数のノード数に対する比，縦軸は変動係数の値を表している．

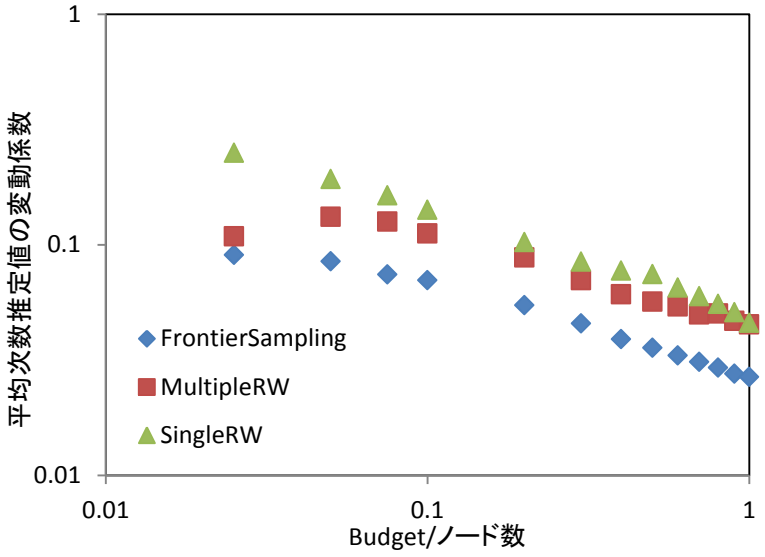

図 5.18　平均次数の変動係数（$n = -1$：低次数ノード優先）

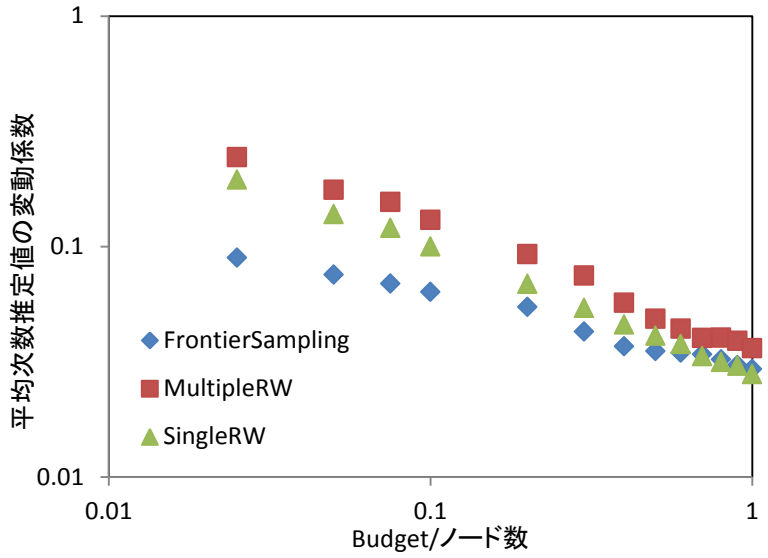

図 5.19　平均次数の変動係数（$n = 0$：MH アルゴリズム）

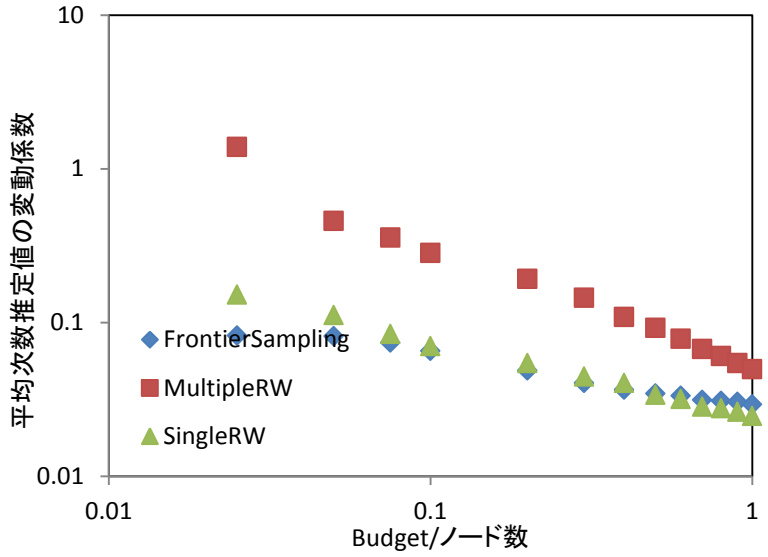

図 5.20　平均次数の変動係数（$n = 1$：古典的ランダムウォーク）

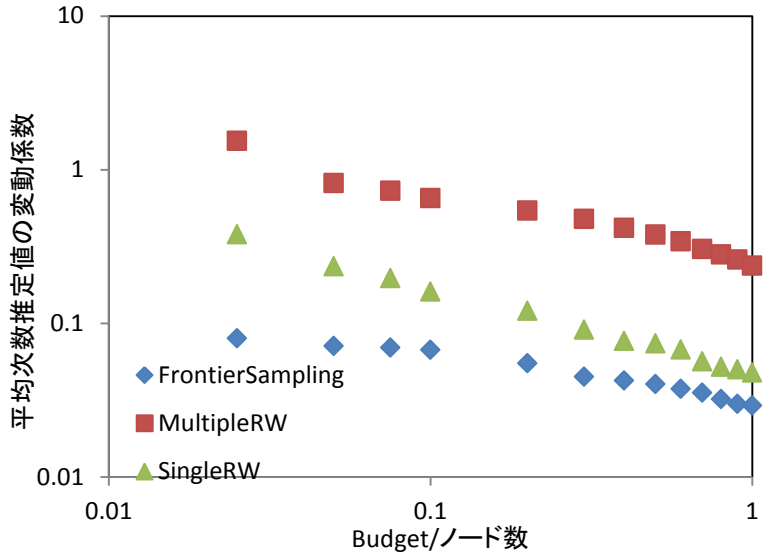

図 5.21　平均次数の変動係数（$n = 2$：高次数ノード優先）

図 5.18～図 5.21 に示されているように，パラメータが-1，2 のとき，Frontier Sampling の変動係数は他のサンプリング手法に比べて小さい．パラメータが 0，1 のとき，Multiple ランダムウォークサンプリングによる平均次数推定値の変動係数は，他のサンプリング手法に比べて変動係数が大きい．一方，Frontier Sampling と Single ランダムウォークサンプリングの変動係数はほぼ同じであるが，Budget が少なく，十分なステップ数が得られない場合には，Frontier Sampling の方が変動係数は小さいことが見て取れる．なお，Budget が大きくなると Single ランダムウォークサンプリングの方が変動係数は小さくなる．

5.3.3　次数分布の推定

Frontier Sampling，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングの 3 通りのランダムウォークについて初期ノードを 100 通り変えて，低次数ノード比率（次数が 3 以下のノードの比率），および高次数ノード比率（次数が上位 10%値を超えるノードの比率）を推定した結果を示す．高次数ノード比率を調べる際には，予め次数の上位 10%値をオフラインで求めておき，ウォーカーが訪問したノードの次数が上位 10%値を超えている頻度から高次数ノード比率を推定させることとした．一般に，低（高）次数ノード比率を推定する際には，低（高）次数ノードを優先訪問させることが望

ましいと考えられるため，バイアス調整パラメータnは-1 から 2 まで変えて，結果を取得した．サンプリング Budget の総量はノード数の 1 割とした．

低次数ノード比率の推定結果を図 5.22 に示す．Multiple ランダムウォークサンプリングはバイアスパラメータ依存性がある．Multiple ランダムウォークサンプリングはパラメータが 0 を超えて大きくなると推定精度が劣化する．一方，やや意外なことに，Frontier Sampling と Single ランダムウォークサンプリングはバイアスパラメータ依存性がほとんど見られず，推定結果が Multiple ランダムウォークサンプリングよりも安定していた．

参考までに，図 5.23 に低次数ノード比率推定値の変動係数を示す．Single ランダムウォークサンプリングは推定結果にはバイアスパラメータ依存性が見られなかったが，変動係数はバイアスパラメータが高くなるほど減少し，バイアスパラメータが 1 を超えたあたりから上昇した． Multiple ランダムウォークサンプリングは，バイアスパラメータが低くなるほど変動係数が減少した．Frontier Sampling は変動係数にもバイアスパラメータ依存性が見られなかった．

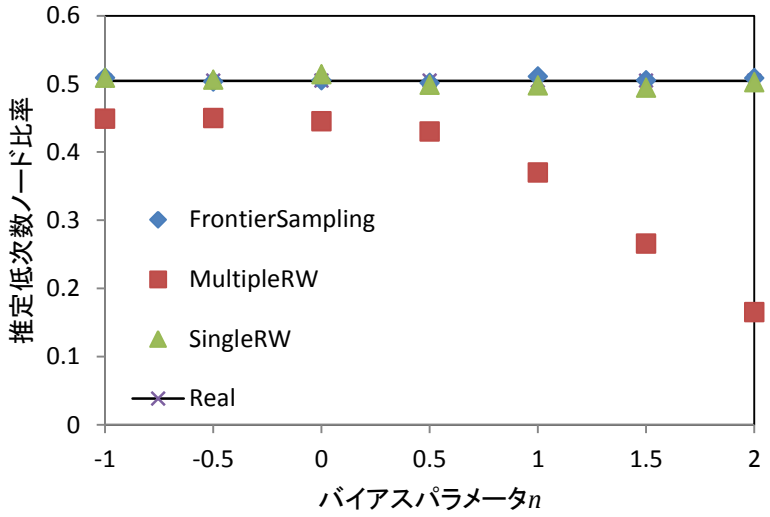

図 5.22　低次数ノード比率推定結果

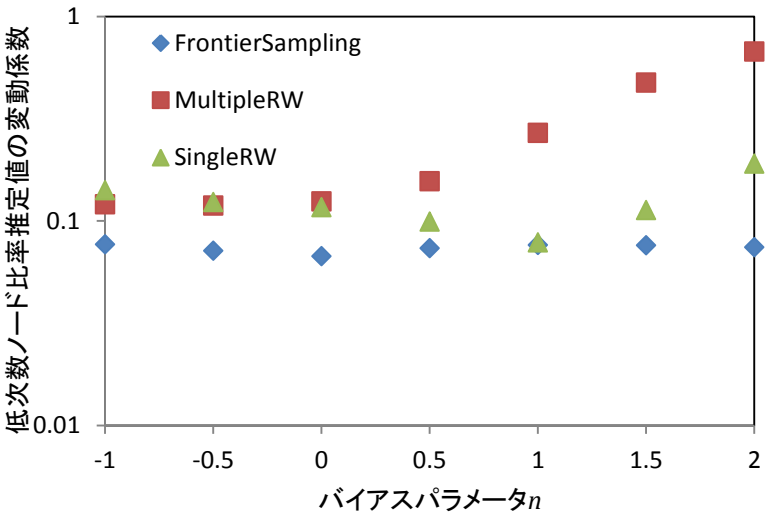

図 5.23　低次数ノード比率推定値の変動係数

次に，高次数ノード比率推定結果をプロットした結果を図 5.24 に示す．やはり Frontier Sampling，Single ランダムウォークサンプリングはバイアスパラメータ依存性がほとんど見られず，推定結果が Multiple ランダムウォークサンプリングよりも安定していた． また，意外なことに，Multiple ランダムウォークサンプリングについては，高次数ノード比率を推定する際にも，バイアスパラメータが低いほど（低次数優先訪問とするほど）推定結果が向上した．

図 5.25 に高次数ノード比率推定値の変動係数を示す．Single ランダムウォークサンプリングは推定結果にはバイアスパラメータ依存性が見られなかったが，変動係数はバイアスパラメータが高くなるほど減少し，バイアスパラメータが 1 を超えたあたりから上昇した． Multiple ランダムウォークサンプリングは，バイアスパラメータが低くなるほど変動係数が減少した．Frontier Sampling は変動係数にもバイアスパラメータ依存性が見られなかった．

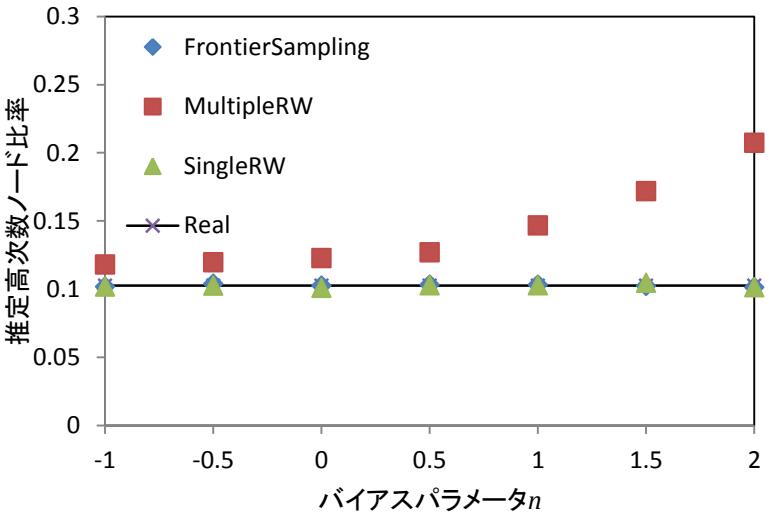

図 5.24　高次数ノード比率推定結果

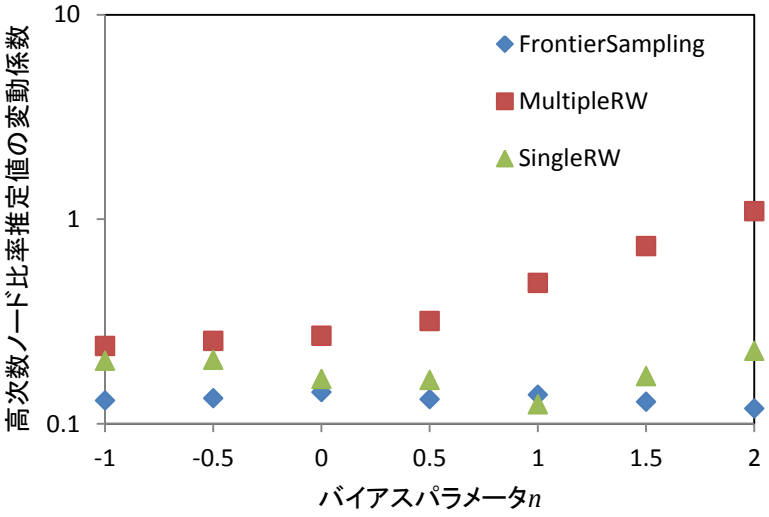

図 5.25　高次数ノード比率推定値の変動係数

5.4　シミュレーション結果：スケールフリーネットワーク

5.4.1　平均次数の推定

ランダムウォークサンプリングでは，ランダムウォークを開始する初期ノードによりサンプリングデータに偏りが発生する．そこで，Frontier Sampling，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングの 3 通りのランダムウォークについて初期ノードを 100 通り変えて，平均次数の推定値を 100 通り求め，その平均を算出した．ノード数が 3000 および 10000 のスケールフリーネットワークで Budget をノード数の 2%からノード数に等しい値まで変えて，推定平均次数をプロットした結果を図 5.26（ノード数 3000）および図 5.27（ノード数 10000）に示す．スケールフリーネットワークでは，バイアス調整パラメータnは 1 に等しいとした．

両方の図からわかるように，Frontier Sampling と Single ランダムウォークサンプリングは正確に平均次数を推定できているが，Multiple ランダムウォークサンプリングは特に Budget が少なく，十分なステップ数が得られない場合，推定誤差が大きいことがわかる．

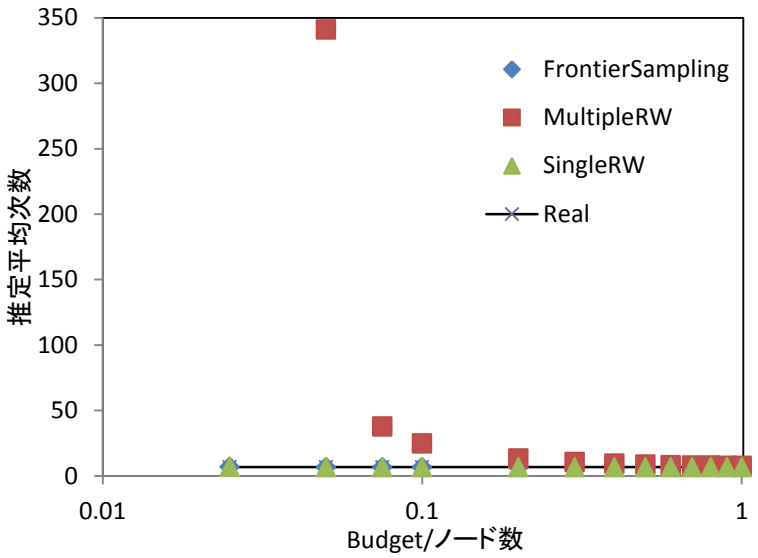

図 5.26　推定平均次数（ノード数　3000）

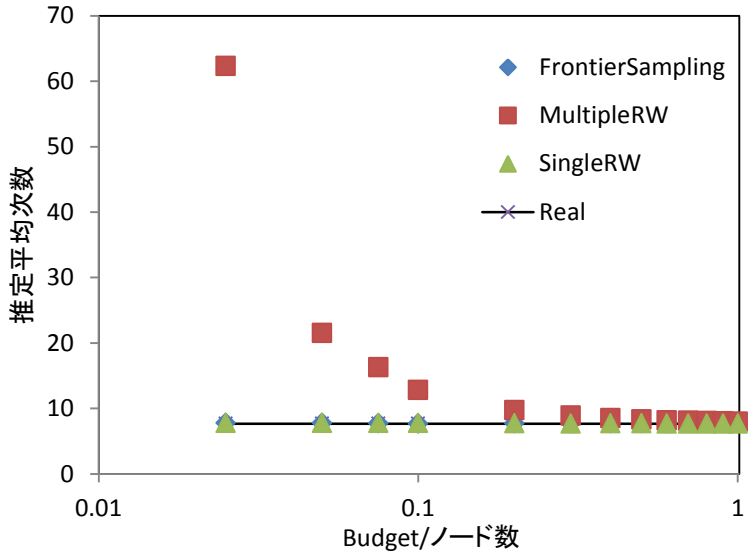

図 5.27　推定平均次数（ノード数　10000）

Multiple ランダムウォークサンプリングの推定誤差は，ウォーカーの初期配置のバイアスに起因していると考えられる．Frontier Sampling では，ウォーカーの初期配置のバイアスが速やかに減少する．Single ランダムウォークサンプリングでは（一つのウォーカーのステップ数を十分大きく取れるので）やはり初期配置バイアスの影響は小さい．

5.4.2　平均次数推定値の変動係数

次に，やはりノード数が 3000 および 10000 のスケールフリーネットワークで Budget をノード数の 2%からノード数に等しい値まで変えて，推定平均次数の変動係数（標準偏差と真の平均次数の比）をプロットした結果を図 5.28（ノード数 3000）および図 5.29（ノード数 10000）に示す．　Multiple ランダムウォークサンプリングによる平均次数推定値の変動係数は，他のサンプリング手法に比べて極めて変動係数が大きい．一方，Frontier Sampling と Single ランダムウォークサンプリングの変動係数はほぼ同じであるが，Budget が少なく，十分なステップ数が得られない場合には，Frontier Sampling の方が変動係数は小さいことが見て取れる．なお，Budget が大きくなると Single ランダムウォークサンプリングの方が変動係数は小さくなる．

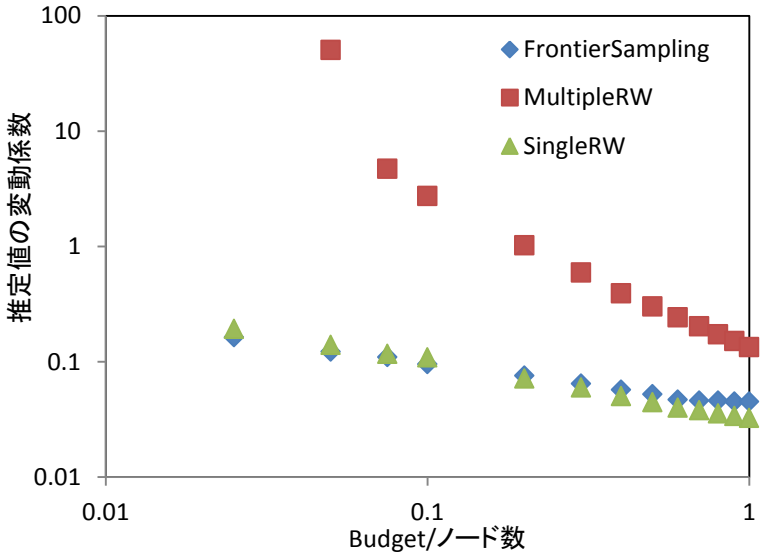

図 5.28 変動係数（ノード数 3000）

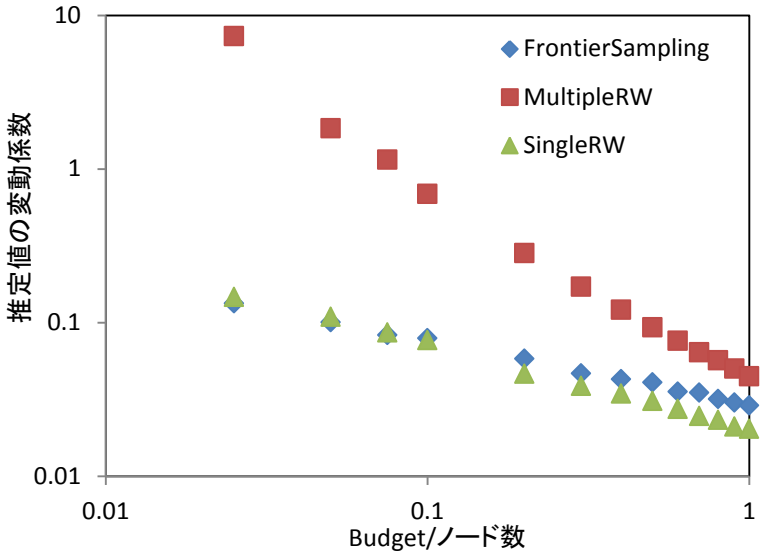

図 5.29 変動係数（ノード数 10000）

5.4.3　次数分布の推定

Frontier Sampling，Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングの 3 通りのランダムウォークについて初期ノードを 100 通り変えて，低次数ノード比率（次数が 3 以下のノードの比率），および高次数ノード比率（次数が上位 10%値を超えるノードの比率）を推定した結果を示す．高次数ノード比率を調べる際には，予め次数の上位 10%値をオフラインで求めておき，ウォーカーが訪問したノードの次数が上位 10%値を超えている頻度から高次数ノード比率を推定させることとした．一般に，低（高）次数ノード比率を推定する際には，低（高）次数ノードを優先訪問させることが望ましいと考えられるため，バイアス調整パラメータnは-1 から 2 まで変えて，結果を取得した．サンプリング Budget の総量はノード数の 1 割とした．

低次数ノード比率の推定結果を図 5.30（ノード数 3000）および図 5.31（ノード数 10000）に示す．Multiple ランダムウォークサンプリング，Single ランダムウォークサンプリングはバイアスパラメータ依存性があり，パラメータが 0 を超えて大きくなると推定精度が劣化する．とくに，Multiple ランダムウォークサンプリングのバイアスパラメータ依存性は顕著である．一方，やや意外なことに，Frontier Sampling はバイアスパラメータ依存性がほとんど見られず，推定結果が Single，Multiple ランダムウォークサンプリングのいずれよりも安定していた．

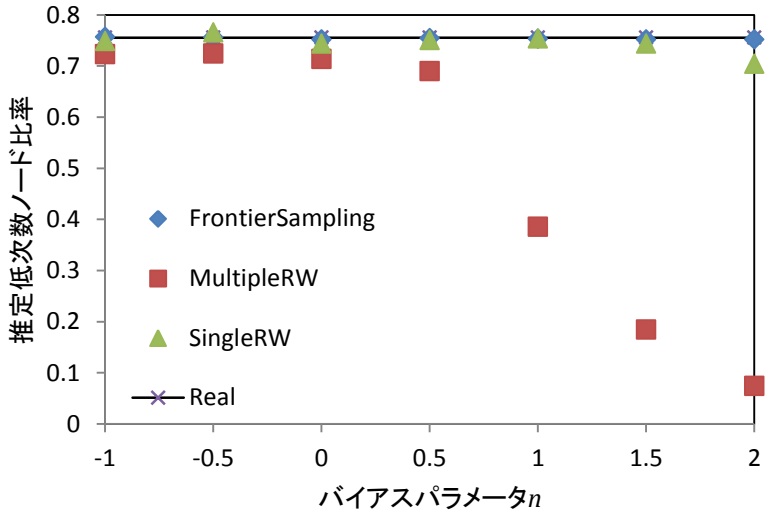

図 5.30 低次数ノード比率推定結果（ノード数 3000）

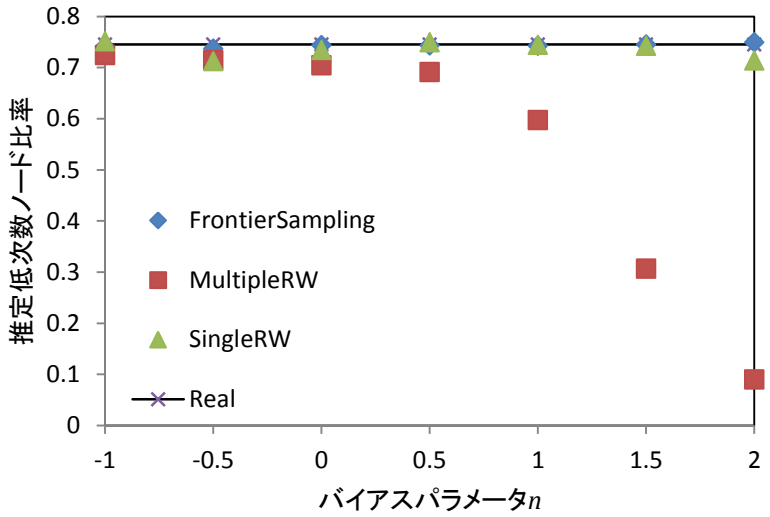

図 5.31 低次数ノード比率推定結果（ノード数 10000）

次に，高次数ノード比率推定結果をプロットした結果を図 5.32（3000 ノード）および図 5.33（10000 ノードに示す）．やはり Frontier Sampling はバイアスパラメータ依存性がほとんど見られず，推定結果が Single，Multiple ランダムウォークサンプリングのいずれよりも安定していた． また，意外なことに，Multiple ランダムウォークサンプリングについては，高次数ノード比率を推定する際にも，バイアスパラメータが低いほど（低次数優先訪問とするほど）推定結果が向上した．

参考までに，図 5.34 に高次数ノード比率推定値の変動係数を示す．Single ランダムウォークサンプリングは推定結果にはバイアスパラメータ依存性が見られなかったが，変動係数はバイアスパラメータが高くなるほど減少し，バイアスパラメータが 1 を超えたあたりから上昇した． Multiple ランダムウォークサンプリングは，バイアスパラメータが低くなるほど変動係数が減少した．Frontier Sampling は変動係数にもバイアスパラメータ依存性が見られなかった．

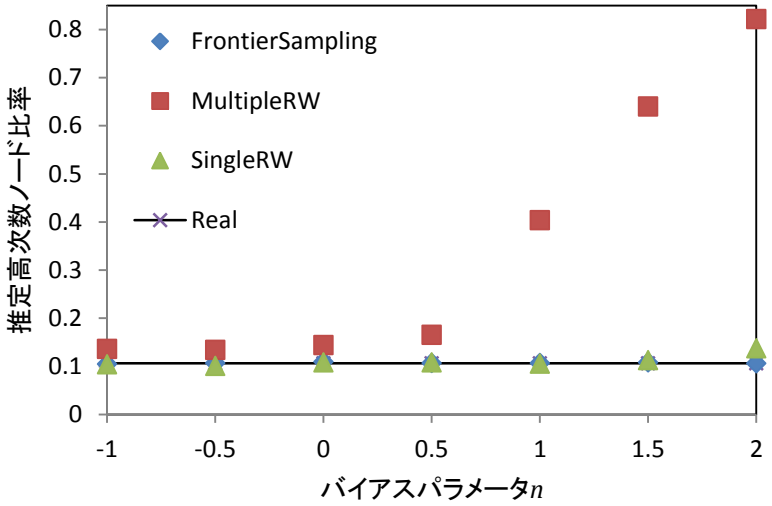

図 5.32 高次数ノード比率推定結果（ノード数 3000）

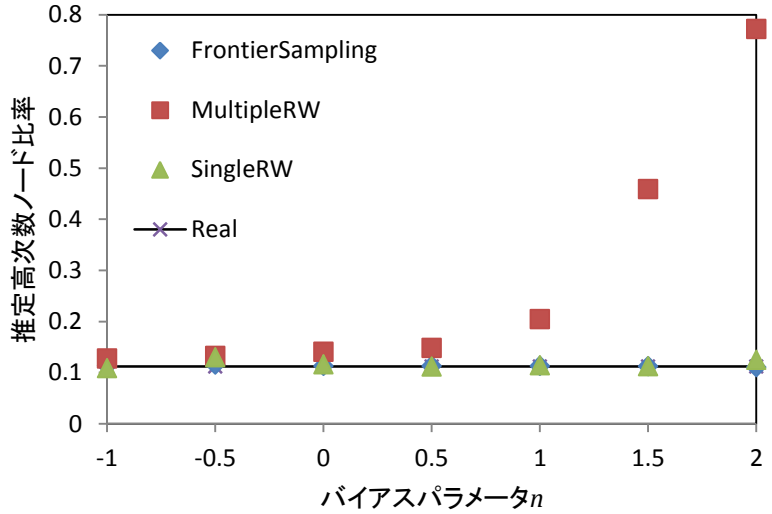

図 5.33 高次数ノード比率推定結果（ノード数 10000）

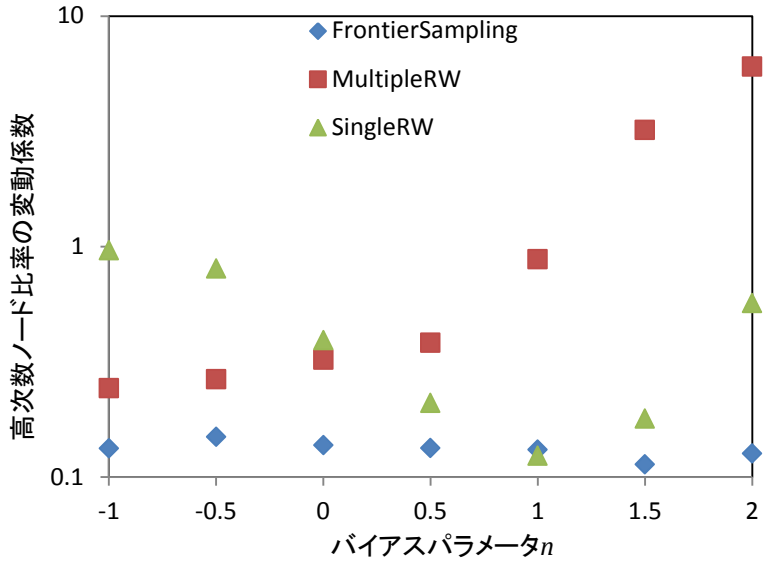

図 5.34 高次数ノード比率推定値の変動係数（ノード数 10000）

第 6 章

結論

本研究では，複数のウォーカーを連携させながら，分析の目的に合わせて，意図的に特定のノード群を高頻度に訪問して情報を収集し，事後に情報に含まれるバイアスを除去して真の特徴量を得るランダムウォークサンプリング手法を提案した．Single ランダムウォークと Multiple ランダムウォークと，サンプリング効率をシミュレーション実験から数値的に比較した．シミュレーション実験にはネットワークのトポロジーデータを用い仮想的に構築したネットワークを利用した．その際，シミュレーションとして，入次数を見ることができない有向グラフを想定し，有向リンクの無向化を行った．シミュレーション結果として平均次数，平均次数の変動係数の推定から，平均次数の真値を推定できる精度の高いサンプリング手法であることを示した．

また，低次数ノード比率，および高次数ノード比率の推定から，遷移確率に依存せず，次数分布の真値を推定できるサンプリング手法であることを示した．以上から，提案手法の有効性を確認した．

参考文献

[1] M.Salehi and H.Rabiee, “A measurement framework for directed networks,” IEEE J. Sel. Areas Commun., vol.31, no.6, pp. 1007-1016, 2013

[2] W.K. Hastings, “Monte Carlo sampling methods using Markov chains and their applications,” Biometrika, vol.57, no.1, pp.97–109, 1970

[3] B.Ribeiro and D. Towsley, “Estimating and sampling graphs with multidimensional random walks,” in Proc. 10th ACM SIGCOMM Conf. on Internet measurement, Melbourne, Australia, 2010

[4] 晒谷亮輔, 塩田茂雄, "ランダムウォークサンプリングで生じるバイアス除去法の比較," 日本オペレーションズ・リサーチ学会春季研究発表会, 1-D-8, 2014

[5] S. Shioda, "Random-walk-based biased sampling for data collection on communication networks," ACM Performance Evaluation Review, vol. 42, issue 2, 2014

[6] “Stanford large network dataset collection”. http://snap.stanford.edu/data/

[7] 樋口 龍雄, 佐藤 公男, "グラフ理論入門 C 言語によるプログラムと応用問題" 日刊工業新聞社, 1999

[8] 矢久保 考介, "複雑ネットワークとその構造" 共立出版, 2013

[9] 増田 直紀, 今野 紀雄, "複雑ネットワークの科学" 産業図書, 2005

謝辞

本研究を進めるにあたり，ご多忙の中，様々なご指導を頂きました塩田茂雄教授に深く感謝いたします.

　また，様々な協力をしてくださった研究室の皆様に感謝の意を表します．誠にありがとうございました.