RF-007

オブジェクト集合に依存した RNG の特性分析
Characteristic analysis of relative neighbor graph depending object sets

小出 明弘† 外岡 達也† 斉藤 和巳†
Akihiro Koide Tatsuya Tonooka Kazumi Saito
青山 一生‡ 澤田 宏‡ 上田 修功‡
Kazuo Aoyama Hiroshi Sawada Naonori Ueda

1. はじめに

多様な現実の複雑ネットワークに共通する性質としてスモールワールド性は注目を集めている．特に，Milgram の実験 [1] における友人関係ネットワーク上でのメッセージ配信や，スモールワールド性を有する人工ネットワークの構成法の一つを数理モデルとして定式化した Watts-Strogatz の研究 [2] が代表的である．そのネットワーク構造面での特徴の一つは，ネットワーク上のリンクを辿り，どのノードからも別のノード（ターゲット）へ，比較的短いステップで，多くの場合に 6（six degrees of separation）程度で到達するパスが存在する点である．一方，ネットワーク機能面での特徴の一つは，Milgram の実験 [1] でも見られるように，ネットワーク上でのターゲットの位置を明確に知らない各ノードが，その近傍の局所情報のみを利用するだけで，ほぼ最短に近いパスを見つけ出し，効率の良いメッセージ配信（ルーティング）を実現できる点である．

このような機能面に着目した研究として，Kleinberg [3] は，ユークリッド空間において格子状に配置したノードの最近傍を結合したネットワークに対して，ある確率モデルに基づきリンク群を付与して構成するネットワークを，Watts-Dodds-Newman [4] は，人物間の類似度に基づく社会距離（厳密には三角不等式を満たさない非類似度）の導入により構成する階層的人間関係ネットワークを対象とした．このようなネットワークを構成すれば，局所情報に基づく効率の良いルーティングが実現できることを理論的に示した．しかしながら，これらの研究では，理論的な実現可能性の追求に主眼が置かれ，実問題を対象とした応用でのスモールワールド性の有効利用について触れられず，応用可能性の追求は重要な研究課題として残されている．

本論文では，工学的な視点に基づき，メッセージ配信をオブジェクトの類似探索問題と捉える．その効率の良い探索を実現する有望なネットワーク構成法として，相対近傍グラフ RNG (Relative Neighborhood Graph) [5] に着目する．具体的には，オブジェクトの総数や、その分布に依存して，構成する RNG 構造がどのように変わるか調査する．実験では，現実データとしては 10 年間分の新聞記事集合を，人工データとしては多次元の一様分布とガウス分布に基づき生成したベクトルをオブジェクト集合として評価する．代表的なネットワーク特徴量に基づく評価結果より，数万単語の新聞記事集合から構成される RNG が，僅か 10 次元程度の人工データのベクトル集合から構成される RNG と類似した性質を持つことを示す．

†静岡県立大学 University of Shizuoka
‡NTT コミュニケーション科学基礎研究所

2. ネットワーク構成法と評価データ

2.1 相対近傍グラフ RNG

オブジェクト集合を $\{p_1, p_2, \ldots, p_N\}$ とする．ここでは，図 1-1 や 1-2 に示すように，2 次元平面上のベクトルを例に説明するので，各オブジェクトを単に点と呼ぶ．一方，任意の 2 点間の距離として通常のユークリッド距離 $d(p_i, p_j)$ を採用する．いま，任意の 2 点を p_i, p_j に対して，2 点間の距離を半径とする円を，点 p_i と p_j のそれぞれを中心として描き，その重なった部分を Lune と呼ぶ．このとき，Lune にその他の点が存在しない場合に，2 点 p_i と p_j の間にリンクを生成して構成したネットワークの構造を相対近傍グラフ RNG (Relative Neighborhood Graph) と呼ぶ．これは，2 点 p_i と p_j が相対近傍であること，すなわち，全ての $p_k \neq p_i, p_j$ について次式が成立するときリンクが生成される．

$$d(p_i,p_j) \le \max_{k \neq i,j} [d(p_i,p_k),(p_j,p_k)] \qquad (1)$$

図 1-1 と 1-2 には，RNG において 2 点 p_i と p_j の間にリンクが生成されない場合とされる場合の例をそれぞれ示す．2 点 p_i と p_j の間でのメッセージ配信を考えれば，図 1-1 の場合には，一方から他方に近づくように点 p_k を経由して配信可能となる．これに対して図 1-2 の場合において，2 点 p_i と p_j の間にリンクが存在しなければ，双方の点から遠ざかる点を経由しなければならない．よって，RNG にはメッセージ配信のためのネットワーク構造として望ましい性質を持つと考えられる．

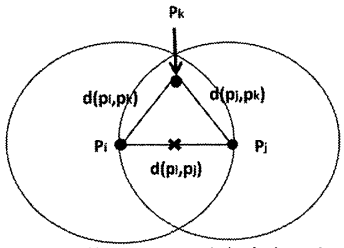

図 1-1 p_i と p_j 間にリンクが生成されない場合

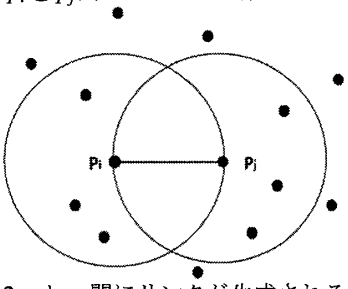

図 1-2 p_i と p_j 間にリンクが生成される場合

2.2 評価データの詳細

本論文では，3 種のデータに RNG を用いて作成されたネットワークを分析する．

一つ目の評価データとして，1992 年 1 月から 2002 年 1 月までの 10 年間分の毎日新聞国際面記事を用いた．総記事数は 64,585 であり，一日当たりの平均記事数はおよそ 15 である．文書データの前処理には，Chasen による形態素解析を施し，助詞などを削除したところ，出現した異なる単語総数は 51,030 となった．各文書は，出現した単語の頻度ベクトルを tf-idf 変換した特徴ベクトルを用いた．すなわち 51,030 次元のベクトルで各文書を表現した．なお，文書の平均単語数は 121.7 であった．一方，文書間の類似度としては，特徴ベクトルのコサイン類似度を用いた．

二つ目の評価データは，予め定めた領域内において指定した数のベクトルをランダムに生成して得られる一様ランダムデータである．ベクトル生成の領域は $[-1, 1]^D$ とした．ここで，D は次元数を表す．

三つ目の評価データは，二つ目の一様ランダムデータと同様にして，ガウス分布を用いてベクトルを生成して得られるガウス分布データである．D 次元ガウス分布の平均は原点 0 に設定し，各軸の分散は 1 で共分散は全て 0 とした．

3. 評価特徴量

本実験で用いる代表的なネットワーク特徴量について述べる．ネットワーク(グラフ) $G=(V_G, E_G)$ をノード(頂点)集合 V_G とリンク(辺)集合 E_G で定義する．ここで，ノード集合を $V_G=\{1, \ldots, N\}$ で，リンク集合を $E_G=\{e_1, \ldots, e_M\}$ で表し，$e_m=\{i; j\} \subset V_G$ かつ $i \neq j$ とする．これは，本論文で対象とする RNG が自己リンクなしの無向グラフになるからである．ネットワーク G において，ノード i が隣接するノード集合を以下で表す．

$$F_G(i)=\{j:\{i,j\}\in E_G\}, \tag{2}$$

3.1 平均次数，平均ノード間距離，ダイアミター

ネットワークの中でそれぞれのノードがいくつのノードと直接つながっているのかをノード次数と呼び，ノード次数をネットワーク全体で平均したのが平均次数である．ノード間距離とは，任意のノード間の最短パス長で定義される．平均ノード間距離とは，全てのノード間についてノード間距離を平均したもので，ノード間の近接性を示す．

ダイアミターとは，任意のノード間の最短パス長の最大値である．ダイアミターが大きくなればなるほど，探索が困難になる．

3.2 k-core

ネットワーク G の k-core とは，G のサブネットワークであって，各ノードが $k-1$ 個以上の隣接ノードをそのサブネットワーク内に持つ場合をいう．より詳細には，与えられた整数 k に対して，以下の条件を満たすノード集合 $V_{C(k)} \subset V_G$ と，リンク集合 $E_{C(k)} \subset E_G$ から構成されるネットワーク $C(k)=(V_{C(k)}, E_{C(k)})$ のことである．

$$V_{C(k)}=\{i:|F_{C(k)}(i)|\geq k-1\}, \tag{4}$$

$$E_{C(k)}=\{em:em\subset V_{C(k)}\}. \tag{5}$$

ここで，$|A|$ は集合 A の要素数を表す．以下では，ノード数が最大となる $C(k)$ を対象とし，その各連結成分 $C^s(k)$ ($1 \leq s \leq S_{C(k)}$) を k-core コミュニティと呼ぶ．ここで，$S_{C(k)}$ は $C(k)$ の連結成分の個数(コミュニティ数) を表す．

4. 分析結果

本章では，3 種のデータのそれぞれで，代表的なネットワークの特徴量を用いてオブジェクト数を変化させて得られる RNG に関する性質を調査する．

4.1 実データの分析結果

実データの分析結果を図 2-1，2-2 に示す．図 2-1 では，横軸にデータの記事数，縦軸に各評価の値を設定しプロットしている．図 2-2 では，横軸に k-core の k の値，縦軸に全ノード数に対する $V_{C(k)}$ に含まれるノード数の比を示している．すなわち，$|V_{C(k)}| / |V_G|$ をプロットしている．

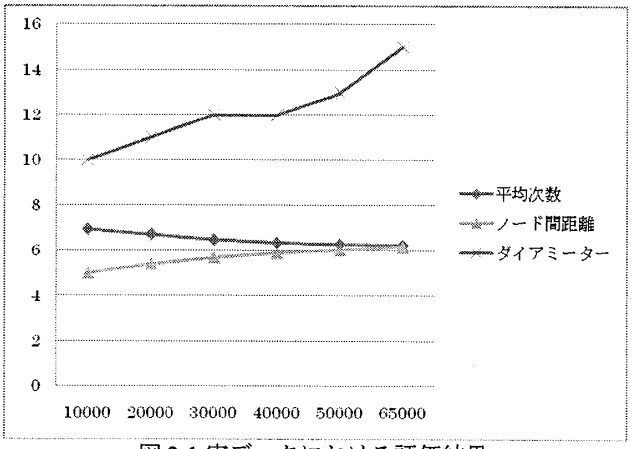

図 2-1 実データにおける評価結果

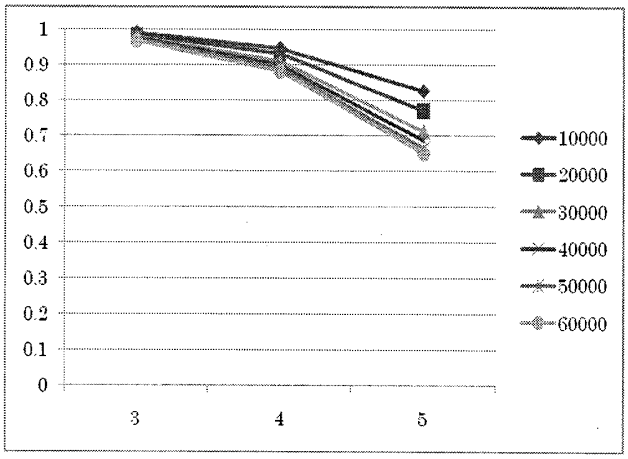

図 2-2 実データにおける k-core 分析結果

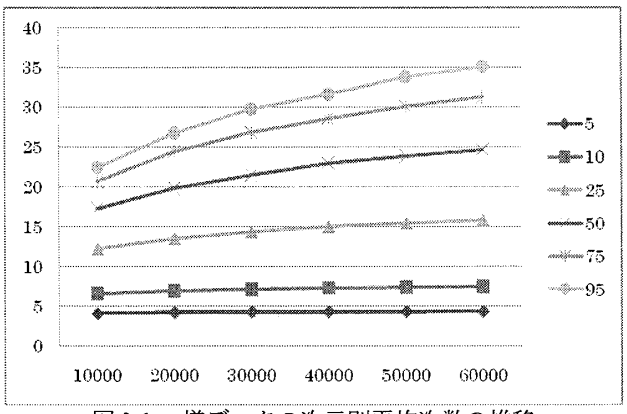

図 3-1 一様データの次元別平均次数の推移

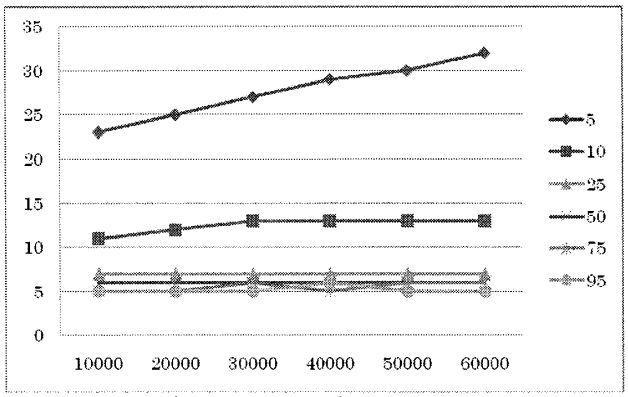

図 3-3 一様データの次元別ダイアミターの推移

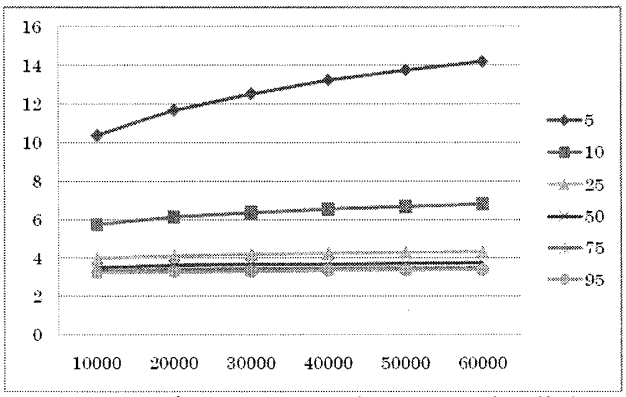

図 3-2 一様データの次元別平均ノード間距離の推移

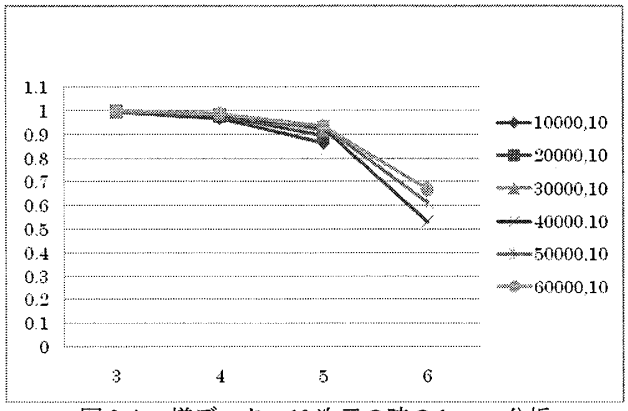

図 3-4 一様データ，10 次元の時の k-core 分析

図 2-1 では，記事数が大きくなるとダイアミターが急激に大きくなっている．一方で，平均次数はやや減少し，平均ノード間距離はやや増加している．また，どちらも記事数の増大に伴い変化量が小さくなっていることが分かる．これらの結果から，実データでは記事数を増加させるとグラフのダイアミターは増加するが，平均次数や平均ノード間距離はそれほど増減しないことが示唆される．

図 2-2 では，記事数が大きくなっても k=5 以上のノードが存在しないことが分かる．このことから，今回利用した実データでは，記事数が増加しても k-core の k が大きくなることはないと推測できる．このことより，RNG においては，k-core で定義されるコミュニティ構造（リンクが密結合する部分）が存在しないことが示唆される．

4.2 一様データの分析結果

一様データの分析結果を図 3-1 から 3-4 にそれぞれ示す．一様データでは，次元数を 5，10，25，50，75，95 と設定し，分析結果を各評価別にプロットした．図 3-1 の平均次数の推移では，そのほかの評価と比較すると規則的に増加していることがわかる．一方，図 3-2，3-3 の平均ノード間期距離，ダイアミターでは，次元数が 5 の時には増加傾向にあるが，次元数が大きくなればなるほどほぼ横ばいの推移になっている．ここで，各次元数の値と図 2-1 で示されている値を比較してみると，実データにおける各評価の値は，一様ランダムデータの 10 次元の値とかなり近い値をとることが示唆された．

次に，図 3-4 の 10 次元の時の k-core 分析と図 2-2 の実データの k-core を比較する．ここで 10 次元の一様データとしたのは，k-core 以外の評価において 10 次元の一様ランダムデータが実データと近い値をとったからである．比較してみたところ，一様データでは，次元数が大きくなるとそれに伴って k が大きくなり，次元数 10 では k=6 まで達することがわかる．さらに次元を高くしてみたところ，k は増加していくことがわかった．

このことから，実データと一様ランダムデータの比較では，実データと次元数 10 の一様ランダムデータが，平均次数，平均ノード間距離，ダイアミターで近い値をとること，また，k-core では実データが記事数が大きくなっても最大の k が変わらないのに対し，一様ランダムデータでは次元数が大きくなると最大の k が増大する傾向にあることが推測できた．

4.3 ガウス分布データの分析結果

ガウス分布データの分析結果を図 4-1 から 4-4 にそれぞれ示す．ガウス分布データにおいても一様データ同様，次元数を 5，10，25，50，75，95 と設定し，分析結果を各評価別にプロットした．一様データの分析結果と比較してみると，平均次数においては，5~10 次元ではお互いに似たような推移をしているが，25 次元まで次元数が大きくなると，ガウス分布データは一様ランダムデータほど値が増加しないことが読み取れる．95 次元で比較してみると，ガウス分布データでは記事数が 10000 から 60000 に増える間に平均次数が約 5 増加するのに対し，一様ランダムデ

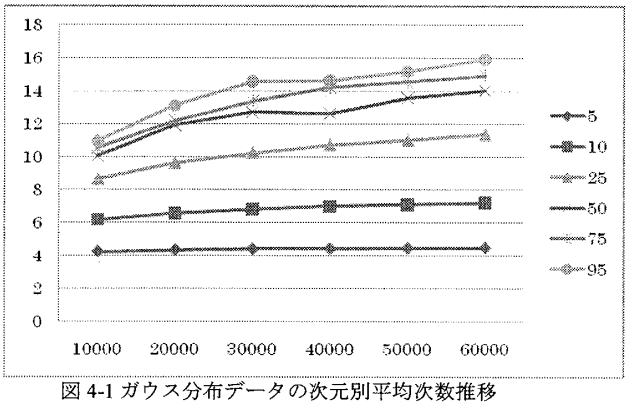

図 4-1 ガウス分布データの次元別平均次数推移

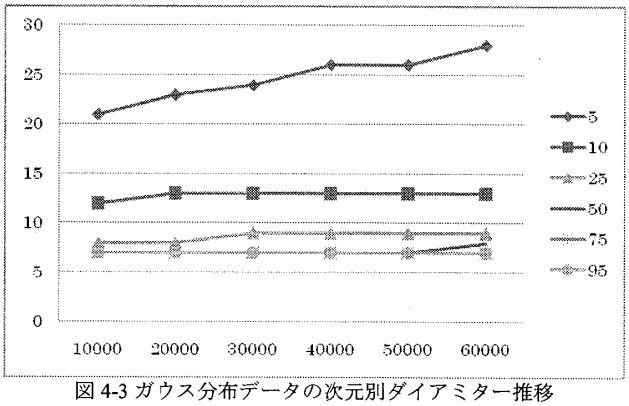

図 4-3 ガウス分布データの次元別ダイアミター推移

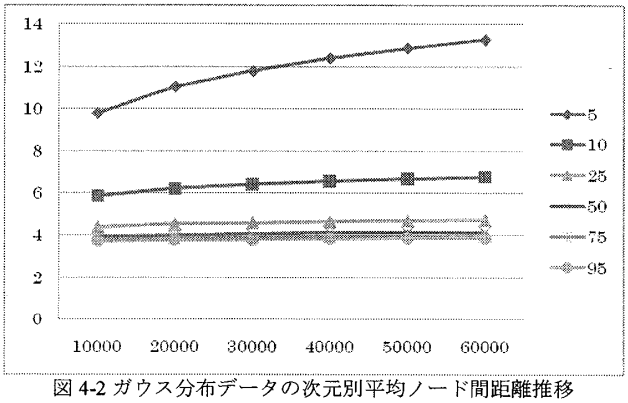

図 4-2 ガウス分布データの次元別平均ノード間距離推移

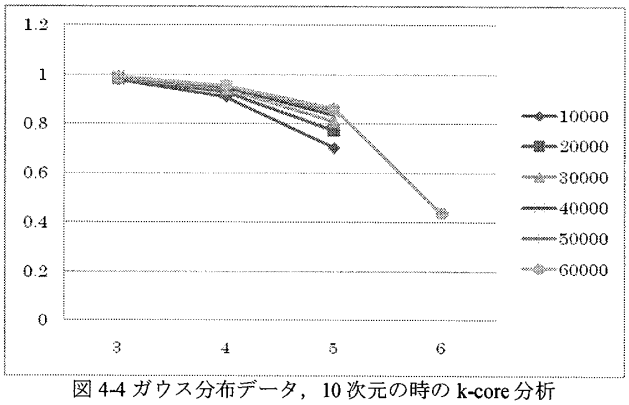

図 4-4 ガウス分布データ，10 次元の時の k-core 分析

ータでは約 15 増加している．一方，平均ノード間距離で比較してみると，どの次元，ノード数を比較してみてもほとんど同じ値を示していることが読み取れる．続いて実データと比較してみると，一様ランダムデータ同様実データにおける各評価の値は，ガウス分布データの 10 次元の値とかなり近い値をとることが示唆された．

次に，図 4-4 の 10 次元の時の k-core 分析と図 2-2 の実データの k-core を比較する．ここで 10 次元のガウス分布データとしたのは，一様ランダムデータ同様 k-core 以外の評価において 10 次元の一様ランダムデータが実データと近い値をとったからである．実データと比較してみると，やはり一様ランダムデータと同様に最大の k が 5 までの実データに対し，10 次元のガウス分布データは k=6 まで存在していることが読み取れる．一様ランダムデータと比較すると，最大の k はお互いに次元数の増加に伴って増加する傾向にあるが，ガウス分布の場合は一様ランダムデータほどは増加しないという結果が得られた．

これら 3 つの異なるデータを用いた分析結果により，平均次数，平均ノード間距離，ダイアミターにおいて実データの値とランダムデータの 10 次元の評価値が似た数値を示すこと，k-core 分析では，実データはノード数が増加しても最大の k は増加しないのに対し，ランダムデータではノード数，次元数の増加に伴って最大の k も増加することが示唆された．また，一様ランダムデータとガウス分布の比較では，平均次数，平均ノード間距離，ダイアミター，k-core のどの評価値で比較してもこの二つのデータは同じような傾向にあると推測できた．また，増加量に関しては一様ランダムデータの方がガウス分布データに比べて増加量が大きいことも読みとることができた．

5. おわりに

本分析では，RNG を用いて作成された異なる 3 つのネットワークデータに対していくつかの評価法を適用し，各ネットワーク間の特徴分析を行った．その結果，実データとランダムデータ間，また，一様ランダムグラフとガウス分布グラフ間にあるいくつかの特徴をみつけることができた．今後は，今回の特徴が現れた要因についての考察を行う．また，今回使用したデータ以外のデータではどのような特徴が見られるのか分析を行う．

謝辞　本研究の一部は科研費(20500109) の助成を受けた．

参考文献

[1]Milgram, S.:The small world problem.*Psychology Today 2*, 60--67 (1967)

[2]Watts, D. J., Strogatz, S. H.:Collective dynamics of `small-world' networks.*Nature 393*, 440--442 (1998)

[3]Kleinberg, J.: Complex networks and decentralized search algorithms.*Proc. Int'l.* Congress of Mathematicians, (2006)

[4] Watts, D. J., Dodds, P. S., Newman, M. E. J.:Identity and search in social networks.*Science 296*, 1302--1305 (2002)

[5] Supowit , J. K. The relative neighborhood graph, with an application to minimum spanning trees. Journal of the ACM (JACM) 30, 428—448 (1983)