実ネットワークに対する，各種リンク予測手法の適性について

Experimental analysis of the applicabilities of link prediction methods for real networks

元田　剛史　　村田　剛志　　佐藤　泰介
Takeshi Motoda　Tsuyoshi Murata　Taisuke Sato

東京工業大学　大学院情報理工学研究科　計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Recently, network analysis has been intensively investigated in several fields of science. Link prediction is the problem of predicting the existence of a link between two entities based on observed links, and it is one of the popular link mining tasks. Although many link prediction methods have been proposed, they have their merits and demerits. In order to obtain the strategies of selecting the best link prediction methods, we perform experiments of six link prediction methods (Common Neighbors(CN), Jaccard's Coefficient(JC), Adamic/Adar(AA), Shortest Path(SP), Preferential Attachment(PA) and Hierarchical Random Graph(HRG)) for 39 real networks. As a result, CN, JC and AA achieve good performance for the networks whose clustering coefficients are more than 0.4. SP achieves good performance for the networks whose average shortest path lengths are more than 3. PA underperforms a random predictor for the networks whose variances of degrees are less than 5. HRG performs consistently well.

1. はじめに

近年，科学の様々な分野においてネットワークで表現されるデータの解析が盛んに行われている [5]．データの解析の一つに，ネットワークの観測部分から未観測部分の構造を予測するリンク予測問題がある [7]．リンク予測問題は，タンパク質間の未観測な相互作用の予測やソーシャルネットワークにおける新たな人間関係の予測など，様々な研究やサービスの発展に寄与するものと期待されている．リンク予測問題に対し様々なリンク予測手法が提案されているが，常に最高の予測精度を達成する手法は知られておらず [6]，現状では高精度なリンク予測の為にネットワークごとに最適な手法の選択が必要である．本研究では 6 種のリンク予測手法 (Common Neighbors(CN), Jaccard's Coefficient(JC), Adamic/Adar(AA), Shortest Path(SP), Preferential Attachment(PA), Hierarchical Random Graph(HRG)) について，Web 上から集めた 39 種のネットワークに対するリンク予測精度を計測し，各種リンク予測手法の適用範囲を調べた．ここではリンク予測問題を，ノードペアをリンクのありそうなものから降順に並べるランキング問題とし，ランキングより算出される AUC(Area Under ROC Curve) をリンク予測精度とする．実験の結果，CN,JC,AA はクラスタ係数 0.4 以上のネットワークに対して高い予測精度を達成し，クラスタ係数 0.1 以下のネットワークに対してのリンク予測精度は低かった．

SP は平均最短経路長が 3 以上のネットワークに対して高い予測精度を達成し，平均最短経路長 3 以下のネットワークに対しては非常に予測精度が低かった．PA は次数の分散が 5 以下のネットワークに対して予測精度が低かった．HRG はどのネットワークにも安定して高い予測精度を達成したが，特に他手法が苦手とするクラスタ係数が小さく次数の分散が小さいネットワーク，或いはクラスタ係数が小さく平均最短経路長が短いネットワークに対して他手法より優れて高い予測精度を達成した．

連絡先: 元田剛史，東京工業大学，motoda@ai.cs.titech.ac.jp

2. リンク予測手法

Getoor らによると，リンク予測問題へのアプローチは 2 種類に大別できる [7]．一方は，Liben-Nowell らのノードペアの類似度を用いたリンク予測手法 [5] のようにネットワークの構造的性質を用いる方法であり，他方はリンク予測に属性情報を用いる方法である．しかし何れを用いても高い精度でのリンク予測は難しい．そこでリンク予測の精度を改善するため，各リンクの有無をノードペアごと独立に予測するのではなく，観測されたネットワークからネットワーク全体のリンク構造を予測する方法が考案された．Clauset らは任意の階層構造を持つネットワークを生成できる Hierarchical Random Graph(HRG) と呼ばれる汎用モデルを定義し，HRG を用いたリンク予測手法を提案している [1]．HRG はパラメータを調整することにより観測部分のネットワークの生成モデルを定義し，その生成モデルを用いて未観測部分のリンクを予測する．本論文では類似度を用いたリンク予測手法 5 種 [5] 及び HRG を用いたリンク予測手法 [1] のリンク予測精度を比較した．以下に各リンク予測手法を紹介する．

2.1 類似度を用いたリンク予測手法 [5]

Liben-Nowell らはネットワークの構造的性質からリンク予測を行う方法を用いて，ある時刻における共著ネットワークから将来の共著関係を予測する研究を行っている [5]．彼らはネットワークにおいて類似度の高いノードペア間に新しいリンクが生まれやすいと仮定し，全ノードペアを類似度の大きいものから降順にランキングすることでリンク予測を行っている．ノード x, y の類似度はネットワークの構造的性質に基づき $score(x, y)$ として算出され，論文 [5] では多数の $score(x, y)$ が定義されている．以下にその一部を紹介する．なお，$\Gamma(x)$ はノード x の隣接ノードの集合を表す．

- Common Neighbors(CN)
  CN は，ノード x とノード y の共通隣接ノード数が多ければ多いほど，2 ノード間にリンクが存在する可能性は高いとする指標である．直感的には,「共通の友人が多い二人は友人である可能性が高い」とする指標である．

$$common(x, y) := \mid \Gamma(x) \cap \Gamma(y) \mid \quad (1)$$

- Jaccard's Coefficient(JC)
  JC は，ノード x とノード y が持つ隣接ノードの総数のうち，共通隣接ノードの数の割合が大きければ大きいほど 2 ノード間にリンクが存在する可能性が高いとする指標である．直感的には,「友人の大半が重なる二人は友人である可能性が高い」とする指標である．

$$Jaccard's(x,y) := \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2}$$

- Adamic/Adar(AA)
  AA は重みつきの CN であり，隣接ノードごとに異なった重みが割り当てられる．重みの大きさは隣接ノードの次数に応じて決められ，次数の小さい共通隣接ノードに対してはより大きな重みが割り当てられる．直感的には,「知り合いが少ない人間を共通の友人とする二人は友人である可能性が高い」とする指標である．

$$Adamic/Adar(x,y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \tag{3}$$

- Preferential Attachment(PA)
  PA は，ノード x とノード y，各々の隣接ノードの数が多ければ多いほど 2 ノード間にリンクが存在する可能性が高いとする指標である．直感的には,「多くの友人を持つ人間ほど，新たな人間と友人になる可能性が高い」とする指標である．

$$preferential(x,y) := |\Gamma(x)| \cdot |\Gamma(y)| \tag{4}$$

- Shortest Path(SP)
  SP はノードペアのネットワーク上での距離が近いほど，ノードペア間にリンクが存在する可能性が高いとする指標である．

2.2 HRG を用いたリンク予測手法 [1, 2, 3]

近年，ネットワークの持つ隠れた階層構造に着目した研究が多く行われている [4]．階層構造は単にクラスタリング結果を表現するに留まらず，ネットワークに含まれる多くの構造情報を正確にとらえることが可能である [1]．Clauset らは論文 [1, 2, 3] において詳細な階層構造の概念を定義し，任意の階層構造を持ったネットワークを生成できる Hierarchical Random Graph(HRG) と呼ばれる汎用モデル，及び HRG を用いたリンク予測手法を提案した．

2.2.1 Hierarchical Random Graph の定義 [2, 3]

HRG は任意の階層構造を持ったネットワークを生成するための汎用モデルである．ここで階層構造は，ネットワークのノードをグループに分け，またそのグループをサブグループに分けるという作業を再帰的に，各ノードが各々個別のサブグループに分類されるまで繰り返すことにより得られる構造を指す．この階層構造は以下に定義する樹形図を用いて表現する．n 個のノードを持つネットワークを G とする．

- 樹形図
  樹形図は二分木で表現され，G のノードを表す葉と，複数の葉による階層構造を表す内部ノード D_i を持つ．この時樹形図を $D = \{D_i, D_2, ..., D_{n-1}\}$ と表す．全てのノードペア $\langle x,y \rangle$ は一つの D_i によって関連付けられる．この D_i はノードペア $\langle x,y \rangle$ の共通の祖先の内，樹形図で最下層に位置する内部ノードである．

HRG は前述の定義に基づいた樹形図 D と，その内部ノード D_i に与えた確率値 θ_i で表現される．確率値 θ_i は，内部ノード D_i の子孫の内左側の部分グラフに属するノードと右側の部分グラフに属するノードの間にリンクが存在する確率である．樹形図 D と確率値 $\vec{\theta}$ の 2 つが与えられた時，HRG$H(D,\vec{\theta})$ は図 1 で表現される．

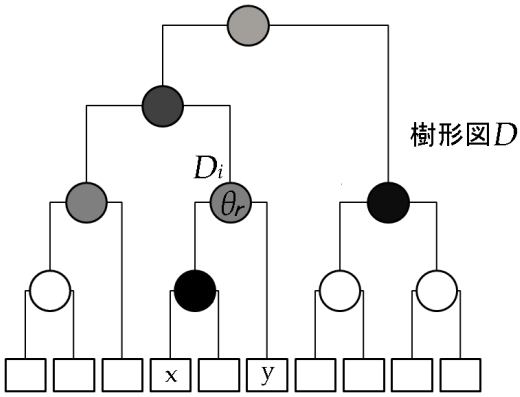

図 1: $H(D,\vec{\theta})$ の一例．□はネットワーク G のノードを表し，○は樹形図 D の内部ノードを表す．内部ノード D_i には確率 θ_r が与えられ，右側の子孫と左側の子孫の間には確率 θ_r でリンクが存在する．本図では確率 1 を白丸，確率 0 を黒丸で表している．$< x,y >$ は x と y の共通の祖先の内，樹形図で最下層に位置する内部ノード D_i によって関連付けられている．

2.2.2 HRG の学習 [2]

本節では，観測したネットワーク G から G を生成したと考えられる $H(D,\vec{\theta})$ を学習する方法について説明する．D が与えられているとき，$H(D,\vec{\theta})$ が G の生成モデルを適切に表現するような $\vec{\theta}$ は最尤法を用いて以下の方法で求められる．
D において，内部ノード D_i によって関連付けられるノードペアの内，G にリンクが存在するノードペアの数を E_i とする．L_i，R_i をそれぞれ，D_i を共通の先祖とする左側部分グラフと右側部分グラフの葉の数とすると，確率モデルとしての HRG の尤度は以下の式で表現できる．

$$L(D,\vec{\theta}) = \prod_{i=1}^{n-1} (\theta_i)^{E_i} (1-\theta_i)^{L_i R_i - E_i} \tag{5}$$

ここで $0^0 = 1$ である．D が固定された場合，$L(D,\vec{\theta})$ を最大化するような確率 $\overline{\theta_i}$ は，各内部ノード D_i に対して以下の式で容易に求められる．

$$\overline{\theta_i} = \frac{E_i}{L_i R_i} \tag{6}$$

これより $\overline{\theta_i}$ が D に対して一意に定まることがわかる．よって尤度を最大とする D を得ることが HRG を学習する際の目標となり，その為に Markov Chain Monte Carlo(MCMC)[2] を用いて D のサンプリングを行う．以下に手順を示す．

1. 内部ノード D_i を，ルートを除く D の内部ノードから一つ，一様にランダムに選択する．

2. 内部ノード D_i に注目すると，D は D_i の部分木及び親を用いて図 2 の A〜C の何れかのパターンで表せる．今

D_i に注目したとき，D の構造が図 2 の A で表せるとすると D の構造は B 又は C のどちらかに組み換え可能である．
D から組み換え可能な二つの構造の内一つを一様にランダムに選択し，組み換え後の樹形図を D' とする．

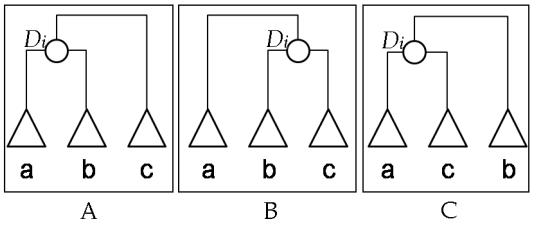

図 2: HRG の各内部ノード D_i(○) は 3 つの部分木 a，b，c(△) と関係を持っている．任意の内部ノード D_i を選んだ時，D は A〜C の何れかのパターンで表現でき，また他のパターンに組み換えが可能である．

3. $D \to D'$ の遷移を採用するか否かを Metropolis-Hastings 規則によって決定する．Metropolis-Hastings 規則とは以下の通りである [2].
もし，$\triangle \log L = \log L(D') - \log L(D)$ が非負ならば $D \to D'$ の遷移を採用する．$\triangle \log L$ が負ならば，確率 $\exp(log\triangle L) = L(D')/L(D)$ で遷移を採用する．

4. 遷移が採用された場合は D' に対して，採用されなかった場合は D に対して，ステップ 1 に戻り同様の処理を行う．

上記手順を繰り返し，高い尤度を持つ $H(D, \vec{\theta})$ をサンプリングする．

2.2.3 HRG を用いたリンク予測 [1]

HRG を用いたリンク予測のアルゴリズムは以下のとおりである．

1. 種となる樹形図をランダムに生成し，マルコフ連鎖を初期化する．

2. MCMC アルゴリズムを HRG の尤度の変化が平衡状態に達するまで走らせる．

3. MCMC によって生成された樹形図から，等間隔で樹形図をサンプリングする．

4. G においてリンクが観測されていない全てのノードペア $\langle i, j \rangle$ に対して，平均確率 $\langle p_{ij} \rangle$ を求める．ここで，$\langle i, j \rangle$ が D_k によって関連付けられるとすると，$\langle p_{ij} \rangle$ はサンプルした各 $H(D, \vec{\theta})$ の θ_k について平均をとることで求められる．

5. $\langle p_{ij} \rangle$ の値に基づいて i と j のノードペアを降順に並べる．

3. 実験

各種リンク予測手法の適用範囲を調べるために，Web 上から集めた 39 種のネットワークのデータに対して実験を行い，リンク予測精度を計測した．

3.1 実験データ

実験には HRG の適用範囲に合わせてノード数 30〜1000 のネットワークを用い，全て無向グラフとして扱った．今回実験に用いたネットワークは多数の分野から網羅的に収集し，その種類は以下の通りである．
人間関係ネットワーク（14 種類），引用・共著ネットワーク（3 種類），辞書ネットワーク（2 種類），生体ネットワーク（5 種類），陸路を表す交通網 (4 種類)，空路を表す交通網 (2 種類)，貿易ネットワーク（3 種類），論理回路ネットワーク（3 種類），食物連鎖ネットワーク（3 種類）．

3.2 実験方法

本研究ではネットワークを入力とし，各ネットワークに対して以下の実験を行った．

1. 入力ネットワークに対してリンクをランダムに 1 割〜9 割取り除いたネットワークを，各割合に対して各々100 個用意する．用意された各々のネットワークはリンク予測問題における観測部分に当たり，取り除かれたリンクが未観測部分に当たる．

2. 用意された各ネットワークに対して，6 種類の手法でリンク予測を行い，得られた各ランキングに対して Area Under ROC Curve(AUC) を計算する．

3. 得られた，各割合 100 個の実験結果に対して，各手法ごとに AUC の平均をとる．各手法ごとの AUC の平均値を本研究では各手法のリンク予測精度と呼ぶ．

4. 与えられたリンクの割合に対して，各手法のリンク予測精度の推移を出力する．

4. 実験結果考察

4.1 CN，JC，AA について

CN，JC，AA はクラスタ係数が 0.4 以上の一部のネットワークに対して，安定して高いリンク予測精度を達成した．クラスタ係数とは距離 2 のノードペアの内リンクが存在するノードペアの割合である．CN，JC，AA は距離 2 のノードペアを上位にランキングするため，クラスタ係数の大きいネットワークに対して今回の結論が得られたものと考えられる．一方，同様の理由でクラスタ係数 0.1 以下のネットワークに対する CN，JC，AA のリンク予測精度は低かった (図 3)．また，クラスタ係数の大きいネットワークの中でも，食物連鎖ネットワークに対するリンク予測精度は，近いクラスタ係数を持つ他のカテゴリのネットワークに対するリンク予測精度より大きく劣っていた．また CN，JC，AA は非連結な部分グラフ間のリンクを予測できないことに加え，距離 2 で離れたノードペア間の共通隣接ノード数に差が生まれないとノードペアをランキングできないことから，観測部分が少ない時は他手法と比べ予測精度が低かった．

4.2 SP について

SP は，平均最短経路長が 3 以上のネットワークに対して高いリンク予測精度を達成した．理由としては以下のことが考えられる．まず，ネットワークでは距離の遠い 2 ノード間にリンク（ショートカット）ができると平均最短経路長が短くなるため，観測したネットワークの平均最短経路長が長い場合，そのネットワークはショートカットができにくい生成モデルから生成されたと考えられる．SP はショートカットのできにくい生成モデルを表現した手法であるため，今回の結果が得られたも

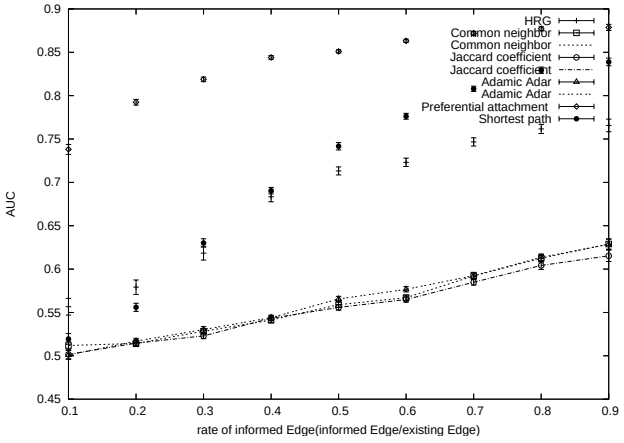

図 3: 生体ネットワーク”E,coli transcription networks”のリンク予測結果．クラスタ係数=0.024．誤差棒は標準誤差を表す．

のと考えられる．一方，ネットワークの平均最短経路長が 3 以下のネットワークに対して予測精度が低かった．これは平均最短経路長が短いネットワークではノードペア間の最短経路長の値に大きな差ができないことが原因の一つと考えられる．

4.3 PA について

PA は，次数の分散が 0〜5 の値を取るネットワークに対してリンク予測精度が低かった．理由としては，PA に従って生成されたネットワークの次数の分散が必ず大きくなることから，次数の分散が小さいネットワークに対して PA の考え方そのものが合致していない可能性が考えられる．図 4 では PA のリンク予測精度は全てのノードペアをランダムに出力した時の値 0.5 を大きく下回っている．この結果からも，先に述べた可能性がリンク予測精度を下げた原因として有力であると考えられる．

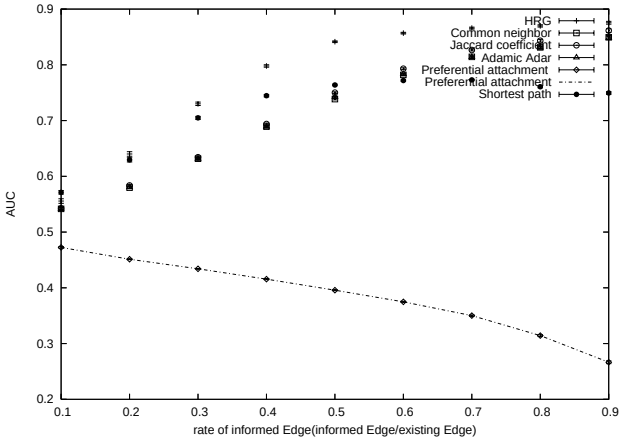

図 4: 人間関係ネットワーク”American College football”のリンク予測結果．次数の分散=0.79．誤差棒は標準誤差を表す．

4.4 HRG について

HRG は特定のネットワーク特徴量に依存することなく，多種のネットワークに対して安定してリンク予測精度が高かった．とりわけ，他手法が苦手とする，クラスタ係数が小さく次数の分散が比較的小さいネットワーク，又はクラスタ係数が小さく平均最短経路長が短いネットワークに対して優れて高い予測精度を達成した．

4.5 各種類ネットワークの特徴値の傾向について

今回用いた 9 種類のカテゴリのネットワークの特徴量に関して以下の傾向が確認された．

カテゴリ名	クラスタ係数	平均最短経路長	次数の分散
人間関係	大		大
引用・共著	大	大	大
辞書	大		大
生体			大
交通網 (陸路)	小	大	小
交通網 (空路)	大	小	大
貿易	大	小	大
論理回路	小	大	小
食物連鎖	大	小	大

表 1: 各カテゴリに属するネットワークの特徴量の傾向を表す．空欄は傾向が確認されなかったことを表す．

5. まとめ

本研究では，様々な分野の実ネットワークに対して網羅的にリンク予測精度を計測することにより，CN，JC，AA，SP，PA，HRG が高い精度を示すネットワークの特徴が確認された．今後の課題として，各手法間のランキング結果の相関等，リンク予測結果のより精細な分析が考えられる．

謝辞

本研究を行うにあたり, 数々のご指導を頂きました亀谷由隆助教に深く感謝いたします．また日常数々のご協力を頂いた佐藤研究室の皆様に感謝いたします．

参考文献

[1] A.Clauset,C.Moore & M.E.J.Newman:Hierarchical structure and the prediction of missing links in networks,Nature,Vol.453,98-101(2008).

[2] A.Clauset,C.Moore & M.E.J.Newman: Hierarchical structure and the prediction of missing links in networks, Supplementary Information, doi:10.1038/nature06830(2008).

[3] A.Clauset,C.Moore & M.E.J.Newman: Structural Inference of Hierarchies in Networks,ICML(2006).

[4] E.Ravasz & A.Barabasi:Hierarchical organization in complex networks. PHYSICAL REVIEW E 67,026112(2003).

[5] D.Liben-Nowell & J.Kleinberg:The Link Prediction Problem for Social Networks,Proc.12th Int.Conf.on Information and Knowledge Management(CIKM),556-559(2004).

[6] 鹿島久嗣:ネットワーク構造予測, 人工知能学会誌,Vol.22,No.3,344-351(2007).

[7] L.Getoor & C.P.Diehl:Link Mining:A Survey,SIGKDD Explorations,Vol.7,No.2,3-12(2005).