1C3-5in

接続行列分解による関係予測

Relation Prediction Using Incidence Matrix Decomposition

横井 祥 *1　梶野 洸 *2　鹿島 久嗣 *3
Sho Yokoi　Hiroshi Kajino　Hisashi Kashima

*1東北大学大学院情報科学研究科システム情報科学専攻
Department of System Information Sciences, Graduate School of Information Science, Tohoku University

*2東京大学大学院情報理工学系研究科数理情報学専攻
Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo

*3京都大学大学院情報学研究科知能情報学専攻
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Sparsity of data sets makes it hard to predict relationships. In this research, we address the sparsity problem to use *incidence matrix decomposition*. We make comparative experiments with the link prediction problem and triadic relation prediction using real data sets and confirm that our method has sparse robustness. Moreover, we give theoretical support in the light of the ratio of the model complexity and the amount of data.

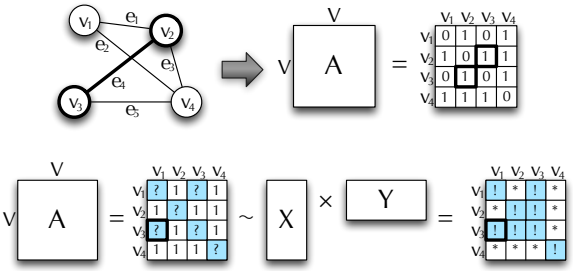

図 1: (従来手法) 隣接行列を用いたグラフのリンク予測—グラフを隣接行列で表現、これを低ランク近似し、復元された隣接行列の (i,j) 成分の大きさを見ることで、頂点対 (v_i, v_j) に将来辺が張られる可能性を予測する。

1. はじめに

生物・化学や社会科学など多くの研究分野において、また SNS など実用的に大規模データを扱う側面において、インスタンス間に特定の 関係 *1 が生じる可能性を予測する問題はしばしば重要なタスクとして現れる [Getoor 05]。関係の予測問題においては 疎性 の問題がしばしば指摘されてきた [Getoor 05, Rattigan 05]。データが疎なとき、つまり可能なインスタンス間の組合わせの数に対して観測されるデータの数が少ないときに学習や予測に困難が生じるという問題である。

関係を予測する問題の典型例として、グラフの リンク予測 問題 [Getoor 05, Liben-Nowell 07]、すなわち時間変化にしたがって辺が増加するグラフにおいて適当な頂点対 (v_i, v_j) に将来辺が張られる可能性の高さを予測する問題が挙げられる。例えば SNS における友人関係の予測はグラフのリンク予測問題と捉えることができる。グラフのリンク予測問題に対しては、グラフの 隣接行列 表現を 低ランク近似 する手法がしばしば用いられる (図 1)。この手法はデータの疎性の問題を抱えている。なぜなら、グラフの表現である隣接行列においては、「頂点対の組合せの数」と「行列の成分の数」、「観測されている辺の数」と「行列中の 1 の数」が一致するからだ。すなわち、データが疎なとき、復元すべき行列の要素数に比して材料となるデータが少なく、学習は困難になる。

連絡先: 横井 祥, 東北大学大学院情報科学研究科システム情報科学専攻, sho.yokoi@gmail.com

*1　下線を引いた語については次章以後で定義を与える。

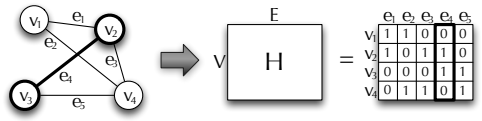

図 2: グラフの接続行列表現—グラフを $|V| \times |E|$ の行列で表現。辺 $e_k = (v_i, v_j)$ と行列の第 k 列目が対応し、その i 行目と j 行目に 1 が現れる。

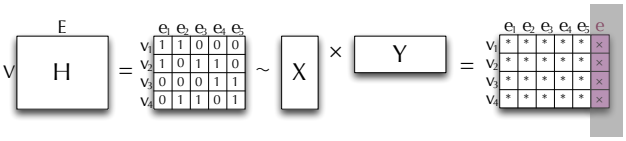

図 3: 接続行列を用いたグラフのリンク予測の非自明性—分解後の行列の中に未観測の頂点対 e の情報が明示的に表れない。

本研究では、グラフのもうひとつの行列表現である 接続行列 を用いることで疎性の問題を解決する。接続行列 とは図 2 のように頂点集合を行側に、辺集合を列側に並べて作られる、グラフの行列表現である。辺 $e_k = (v_i, v_j)$ に対応して (i,k) 成分と (j,k) 成分の値がそれぞれ 1 となる。接続行列によるグラフの行列表現はデータの疎性の問題を緩和していると考えられる。なぜなら、「観測されている辺の数」が行列中の 1 の数として現れる点は隣接行列と同様だが、「頂点対の組合せの数」は行列のサイズに反映されず、データが疎のときは行列のサイズからして小さくなるので、復元すべき行列の要素数に対してその材料となるデータの割合が小さくならない。

接続行列を用いることによって疎性の問題が解決されるとしても、肝心のリンク予測の方法が自明ではない。接続行列を行列分解ないし低ランク近似すること自体は可能ではあるが、近似後の行列には未知の頂点対 $e = (v_l, v_m)$ の情報が明示的に現れず、隣接行列分解時に用いた「行列穴埋め」の考え方からは明らかにならない (図 3)。本研究では、接続行列 H を $H \approx XY$ と分解した後、「頂点対の潜在表現 $\boldsymbol{y}$ をうまくみ

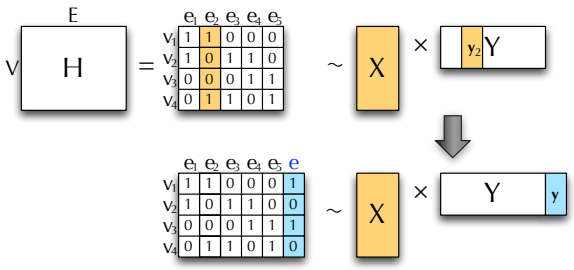

図 4: (提案手法) 接続行列を用いたグラフのリンク予測のアイデア—グラフを接続行列で表現、これを分解する。既存のリンク $\boldsymbol{e}_2$ を表す縦ベクトルは $X\boldsymbol{y}_2$ により作られる。どの既存のリンクも X を通して作られる。うまい $\boldsymbol{y}$ を見つけて、$X\boldsymbol{y}$ が「頂点対 $\boldsymbol{e}$ を表す縦ベクトル」に近ければ頂点対 $\boldsymbol{e}$ にリンクが張られそうだと考える。

つけて、$X\boldsymbol{y}$ で頂点対を表す縦ベクトル (二箇所に 1 が現れる縦ベクトル) を作ることができればリンクが張られやすいと言えるだろう」という考え方に基づきリンク予測をおこなう (図 4)。

なお、一般の多項関係予測問題に対しても、グラフのリンク予測問題とほとんど同様にして提案手法を適用できる。実験ではグラフのリンク予測問題だけでなく三項関係の予測問題も扱った。

以下の章で提案手法の詳細を述べ、提案手法が実験的にも理論的にもデータの疎性に対して頑健であることを示す。

2. 問題設定

PU 学習 (*PU learning: positive unlabeled learning*) 問題とは、有限集合 $\mathcal{X}$ 全体とその部分集合である正例集合 $\mathcal{X}_\mathrm{P}$ が与えられ、未観測 (unlabeled) のデータ $\mathcal{X}_\mathrm{U} = \mathcal{X} \setminus \mathcal{X}_\mathrm{P}$ の正例となる可能性の高さを予測する関数 $f : \mathcal{X} \to \mathbb{R}$ を学習する二値分類問題である。

有限集合の族 $\{X_i\}_{i=1}^n$ に対して、その直積の部分集合 $R \subset (X_1 \times \cdots \times X_n)$ を集合族上の **関係** といい、$(x_1, x_2, \ldots, x_n) \in R$ のとき $(x_1, x_2, \ldots, x_n)$ は **関係を持つ** という。以下、関係の予測問題を PU 学習の枠組みで定義する。

訓練データとして頂点全体と辺の一部が観測されている無向グラフ $G = (V, E_\mathrm{P})$ が与えられたとき、他の頂点対 $E_\mathrm{U} := (V \times V) \setminus E_\mathrm{P}$ に辺が張られる可能性の高さを予測する関数 $f : V \times V \to \mathbb{R}$ を学習する問題をグラフの **リンク予測** 問題 [Popescul 03, Getoor 05, Liben-Nowell 07] という。すなわち、グラフのリンク予測問題を二項関係予測問題として定式化する。

訓練データとして 3 個の有限集合 A, B, C と、関係を持つことが分かっている三つ組の集合 $R_\mathrm{P} = \left\{(a^{(i)}, b^{(i)}, c^{(i)})\right\}_{i=1}^N \subset (A \times B \times C)$ が与えられたとき、他の三つ組 $R_\mathrm{U} = (A \times B \times C) \setminus R_\mathrm{P}$ がそれぞれ関係を持つ可能性の高さを予測する関数 $f : (A \times B \times C) \to \mathbb{R}$ を学習する問題を **三項関係予測** 問題とよぶ。

3. 準備

3.1 行列分解と低ランク近似

行列を複数の行列の積であらわすことを **行列分解** という。与えられた行列をより低いランクの行列の積に分解して元の行列を近似することを行列の **低ランク近似** という。それぞれ目的に応じて様々な分解方法やその数値計算手法が提案されている。今回実験で扱った **特異値分解** (*SVD*) は行列分解手法のひとつで、与えられた実行列 $X \in \mathbb{R}^{m \times n}$ を 3 つの行列 $U \in \mathbb{R}^{m \times k}, \Sigma \in \mathbb{R}^{k \times k}, V \in \mathbb{R}^{n \times k}$ の積 $U\Sigma V^\mathsf{T}$ に分解する。ここで $k := \operatorname{rank} X$。また、特異値を降順に並べた対角行列 Σ について特異値を $k' < k$ 個のみ残して残りを 0 とした新しい行列 $\tilde{\Sigma}_{k'}$ を用い $\tilde{X}_{k'} := U\tilde{\Sigma}_{k'}V^\mathsf{T} \approx X$ とすると、$\tilde{X}_{k'}$ はフロベニウスノルムでの最良近似を与える：

$$\tilde{X}_{k'} = \operatorname*{arg\,min}_{X' \in \mathbb{R}^{m \times n}} \|X - X'\|_\mathrm{F}^2 \text{ s.t. } \operatorname{rank} X' \leq k'$$

この低ランク近似を *truncated SVD* という。

3.2 テンソル分解

与えられたテンソルをより小さなテンソルや行列の積や和の形に分解して元のテンソルを近似することを **テンソル分解** (*tensor decomposition*) という。多項関係予測問題を扱う際、グラフのリンク予測問題の既存手法である隣接行列分解の自然な拡張としてテンソル分解がしばしば用いられる [Kolda 09] 。

今回実験の比較手法で用いた **CP 分解** (*CP: CANDECOMP/PARAFAC*) はテンソル分解手法のひとつで、与えられたテンソル $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I \times J \times K}$ をランク 1 テンソル [*2] の線形和に近似する：

$$\boldsymbol{\mathcal{X}} \approx \tilde{\boldsymbol{\mathcal{X}}} = (\tilde{x}_{ijk}) := \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r$$

ただし $A = (a_{ir}) = (\boldsymbol{a}_r)_{r=1}^R \in \mathbb{R}^{I \times R}$。$B$、$C$ も同様。$\tilde{\boldsymbol{\mathcal{X}}}$ の各要素 $\tilde{x}_{i,j,k}$ は $\tilde{x}_{ijk} = \sum_{r=1}^R a_{ir}b_{jr}c_{kr}$ により計算できる。

4. 提案手法

4.1 概要

提案手法は関係の予測問題に対して次の手順で取り組む。(1) 訓練データの接続行列への変換、(2) 学習：接続行列の分解、(3) 予測：未知データに対するスコアの算出。以下、グラフのリンク予測問題と三項関係予測問題に対する手続きを具体的に示す。

4.2 グラフのリンク予測

4.2.1 訓練データの接続行列への変換

無向グラフ $G = (V, E_\mathrm{P})$ に対して、**接続行列** $H = (h_{i,j}) \in \mathbb{R}^{|V| \times |E_\mathrm{P}|}$ を次のように構成する

$$h_{i,j} = \begin{cases} 1 & (v_i \in e_j, e_j \in E_\mathrm{P}) \\ 0 & (\text{otherwise}) \end{cases}$$

H の各行が頂点 $v \in V$ を、各列が辺 $e \in E_\mathrm{P}$ を表し、各列にはちょうど 2 つの 1 が現れる (図 2)。

4.2.2 学習：接続行列の分解

H を truncated SVD により 2 個の行列に分解する：$H \approx XY$。ただし $X := U\tilde{\Sigma}_k \in \mathbb{R}^{|V| \times k}$、$Y := V^\mathsf{T} \in \mathbb{R}^{k \times |E_\mathrm{P}|}$。

4.2.3 予測：未知データに対するスコアの算出

頂点対 (v_i, v_j) を表す接続行列 H の縦ベクトルを

$$\mathbf{1}_{i,j} := (0, \ldots, 0, \underset{i}{1}, 0, \ldots, 0, \underset{j}{1}, 0, \ldots, 0)^\mathsf{T}$$

*2 ここでは縦ベクトル 3 つの **直積** (*outer product*) $\circ$ で構成されるテンソルをランク 1 テンソルとよび、ランク 1 テンソルいくつ分の線形和で表されるかをもってテンソルの **ランク** とよんでいる。

とおく。「$\boldsymbol{y} \in \mathbb{R}^k$ を動かして $X\boldsymbol{y}$ が最も $\mathbf{1}_{i,j}$ に近づいたときの距離」dist $((v_i, v_j))$ は次のように求まる：

$$\begin{aligned}\mathrm{dist}\left((v_i, v_j)\right) &:= \min_{\boldsymbol{y}\in\mathbb{R}^k} \|X\boldsymbol{y} - \mathbf{1}_{i,j}\|_2^2 \\ &= \| \left(X(X^\mathsf{T}X)^{-1}X^\mathsf{T} - I_{|V|}\right)\mathbf{1}_{i,j}\|_2^2\end{aligned}$$

ただし $I_{|V|} \in \mathbb{R}^{|V|\times|V|}$ は単位行列。ここで

$$W = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{|V|}) := X(X^\mathsf{T}X)^{-1}X^\mathsf{T} - I_{|V|}$$

は i, j に依らず X のみで定まるので、W を予め計算しておけば、dist $((v_i, v_j))$ は $\boldsymbol{w}_i, \boldsymbol{w}_j$ により簡単に求めることができる：

$$\mathrm{dist}\left((v_i, v_j)\right) = \|W\mathbf{1}_{i,j}\|_2^2 = (\boldsymbol{w}_i + \boldsymbol{w}_j)^\mathsf{T}(\boldsymbol{w}_i + \boldsymbol{w}_j)$$

dist $((v_i, v_j))$ の 小ささ をスコアとして、頂点対 (v_i, v_j) に辺が張られる可能性の高さを予測する。

4.3 三項関係予測

訓練データ $A = \{a_1, \ldots, a_{|A|}\}$, $B = \{b_{|A|+1}, \ldots, b_{|A|+|B|}\}$, $C = \{c_{|A|+|B|+1}, \ldots, c_{|A|+|B|+|C|}\}$, $R_\mathrm{P} = \left\{(a^{(i)}, b^{(i)}, c^{(i)})\right\}_{i=1}^{|R_\mathrm{P}|} \subset A \times B \times C$ [*3] に対して接続行列 $H = (h_{i,j}) \in \mathbb{R}^{(|A|+|B|+|C|)\times|R_\mathrm{P}|}$ を次のように構成する：

$$h_{i,j} = \begin{cases} 1 & (a_i \in r_j, r_j \in R_\mathrm{P}) \\ 1 & (b_i \in r_j, r_j \in R_\mathrm{P}) \\ 1 & (c_i \in r_j, r_j \in R_\mathrm{P}) \\ 0 & (\text{otherwise}) \end{cases}$$

接続行列 H は、行方向には集合 A, B, C の要素が添字の順に並び、列方向には関係 $r \in R_\mathrm{P}$ が並ぶ。各列にはちょうど 3 つの 1 が現れる。学習と予測はグラフのリンク予測問題と同様。

5. 実験

グラフのリンク予測問題および三項関係予測問題について、実データにて提案手法の性能を確認した。

5.1 グラフのリンク予測

KONECT データセット [*4] より、辺に重みのない無向グラフを $|V|$ の数が小さいものからカテゴリを問わず 16 種取り上げた。比較手法には隣接行列の Truncated SVD による低ランク近似を用いた。

実験の結果、特にグラフの規模 ($|V|$) が同程度で、その密度 ($|E|$) に大きく差があるグラフ同士を比べると、より疎な ($|E|/|V|^2$ が小さい) グラフほど提案手法が相対的に良好な予測結果を示していることが分かった (表 1)。提案手法の疎性に対する頑健さについては次章で詳しく考察する。

5.2 三項関係予測

Kinships, UMLS, Nations の各データセットを取り上げた。比較手法には、三項関係を三階テンソルに変換し CP 分解する手法を採用した。比較手法の実験結果は [Nickel 11] による。

*3 接続行列への変換の便宜上、$A \cup B \cup C$ 全体で添字がユニークになるように A の要素から順に添字付けをおこなった。

*4 http://konect.uni-koblenz.de/networks/

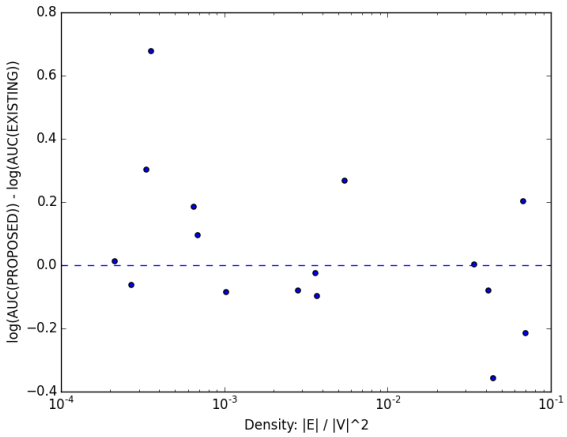

図 5: グラフのリンク予測問題の実験における「データの疎性」と「提案手法の予測結果の優位性」の関係。

実験結果は表 2 の通り。また、実験に用いた三項関係の規模 ($|A|\times|B|\times|C|$) は同程度であり、その予測精度を密度 ($|R|$) の多寡で比較すると、この実験でもより疎な ($|R|/(|A|\times|B|\times|C|)$ が小さい) 訓練データであるほど提案手法が相対的に良好な予測結果を示していることが分かる。

6. 考察

6.1 実験結果に見られる提案手法の疎性に対する頑健性

はじめに、提案手法が疎性に対して頑健であることを実験結果から確かめる。問題が与えられたとき、可能な関係の組合わせの数全体に占める正例集合 R_P の割合 Density(R_P) を訓練データの **密度** 、密度の低さを **疎性**、密度が小さいとき訓練データが **疎** であるとよぶことにする。例えば無向グラフのリンク予測問題 $G_\mathrm{P} = (V, E_\mathrm{P})$ の訓練データの密度 Density(E_P) は $2|E_\mathrm{P}|/|V|^2$、三項関係予測問題 $A, B, C, R_\mathrm{P} \subset A \times B \times C$ の訓練データの密度 Density(R_P) は $|R_\mathrm{P}|/|A| \times |B| \times |C|$ となる。

グラフのリンク予測実験の結果を、横軸をデータの密度、縦軸を提案手法の予測結果の優位性 [*5] として図示したものが (図 5) である。図の左側、つまりグラフが疎である場合ほど提案手法が既存手法に対して良い予測結果を出している傾向が見てとれる。

表 1: グラフのリンク予測問題の実験結果— 既存手法と提案手法の AUC-ROC を比較しその大きい方を太字とした。

#V	#E	Sparsity	EXISTING		PROPOSED
198	2742	DENSE	**0.912**	>	0.736
212	244	**SPARSE**	0.555	<	**0.725**
1858	12534	DENSE	**0.869**	>	0.848
1870	2277	**SPARSE**	0.565	<	**0.680**
2426	16631	DENSE	**0.885**	>	0.816
2888	2981	**SPARSE**	0.506	<	**0.996**
6327	147547	DENSE	**0.979**	>	0.888
6474	13895	**SPARSE**	0.622	<	**0.843**

表 2: 三項関係予測問題の実験結果—既存手法と提案手法の AUC-PR を比較しその大きい方を太字とした。

#A	#B	#C	#R	Sparsity	EXISTING		PROPOSED
104	104	26	10790	DENSE	**0.94**	>	0.926
135	135	49	6752	**SPARSE**	0.95	<	**0.965**
125	125	57	2565	**SPARSE**	0.83	<	**0.996**

*5 提案手法の AUC-ROC / 既存手法の AUC-ROC

表 3: グラフのリンク予測問題における既存手法と提案手法の「モデルの良さ」の比較。

	ModelComplexity	DataSize	ModelEfficiency
E	$k_e \times 2\|V\|$	$2\|E_{\mathrm{P}}\|$	$\dfrac{2\|E_{\mathrm{P}}\|}{k_{\mathrm{E}} \times 2\|V\|}$
P	$k_h \times (\|V\| + \|E_{\mathrm{P}}\|)$	$2\|E_{\mathrm{P}}\|$	$\dfrac{2\|E_{\mathrm{P}}\|}{k_h \times (\|V\| + \|E_{\mathrm{P}}\|)}$

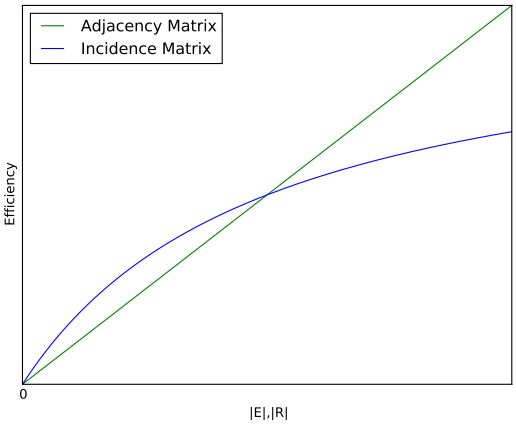

図 6: 関係を予測する問題における「訓練データの密度」と「モデルの良さ」の関係

6.2 モデルの複雑さと訓練データが与える情報の比

なぜ疎なデータに対して提案手法は既存手法に対して相対的に優位となったのだろうか。モデルの複雑さに対して訓練データがどの程度の「情報」を与えるかという観点でモデルを見ると、その理由が見えてくる。モデルを M、解きたい関係予測問題を P とおく。学習時に決定すべきモデルパラメータの数を、問題 P を解く際の **モデル M の複雑さ** $\mathrm{ModelComplexity}(M, P)$ とよぶことにする。すなわち分解後の行列やテンソルの成分の数をモデルの複雑さとする。これはモデルにも問題にも依存する。また、分解前の行列に埋まる 1 の数を問題がモデルに提供する **情報** $\mathrm{DataSize}(M, P)$ とよぶ。これもモデルと問題それぞれに依存する。これらを用いて **モデルの良さ** $\mathrm{ModelEfficiency}(M, P)$「1 モデルパラメータあたりに問題が与える情報」を次のようにおく：

$$\mathrm{ModelEfficiency}(M, P) := \frac{\mathrm{DataSize}(M, P)}{\mathrm{ModelComplexity}(M, P)}$$

たとえばグラフのリンク予測問題における既存手法と提案手法の「モデルの良さ」は表 3 の通り。ここで $k_{\mathrm{E}}, k_{\mathrm{P}}$ はそれぞれ低ランク近似やテンソル分解におけるランクを表す。

グラルのリンク予測問題と三項関係予測問題ともに、$|V|$ すなわちグラフの規模を固定して $|E_{\mathrm{P}}|$ や $|R_{\mathrm{P}}|$ すなわち訓練データの密度を変化させると、既存手法と提案手法の「モデルの良さ」 $\mathrm{ModelEfficiency}(M, G)$ は図 6 のように変化する。特に訓練データが疎なとき (図 6 で左側に行くとき) に提案手法は $\mathrm{ModelEfficiency}(M, G)$ が立ち下がりにくく、したがってデータの疎性に対してより頑健な手法であると考えられる。

ここで定義した「モデルの良さ」は実験結果にもよく当てはまっている。グラフのリンク予測実験について、既存手法の学習モデルを M_{E}、既存手法の学習モデルを M_{P} とし、横軸に提案手法の **モデルの相対的な良さ** $\mathrm{Advantage}(G) := \mathrm{ModelEfficiency}(M_{\mathrm{P}}, G)/\mathrm{ModelEfficiency}(M_{\mathrm{E}}, G)$ を、縦軸

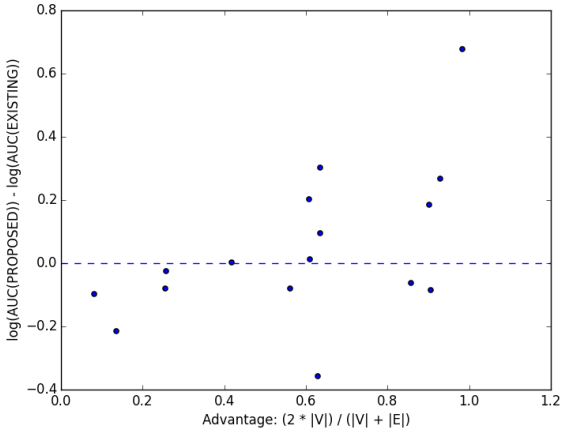

図 7: グラフのリンク予測問題の実験における「提案手法のモデルの相対的な良さ」と「提案手法の予測結果の優位性」の関係。Y 軸は e を底とした対数目盛。

に提案手法の予測精度の相対優位性 [*6] をプロットすると (図 7) のようになり、モデルの相対的な良さ $\mathrm{Advantage}(G)$ という指標が予測結果の相対的な優位性を測りうる指標となっていることが見てとれる。

7. おわりに

接続行列の低ランク近似を用いた新しい関係予測手法を提案した。提案手法が疎性の問題に対して頑健な性質を持つことを実験的に示した。また理論的な考察を与えた。

参考文献

[Getoor 05] Getoor, L. and Diehl, C. P.: Link mining: a survey, *ACM SIGKDD Explorations Newsletter*, Vol. 7, No. 2, pp. 3–12 (2005)

[Kolda 09] Kolda, T. G. and Bader, B. W.: Tensor decompositions and applications, *SIAM Review*, Vol. 51, No. 3, pp. 455–500 (2009)

[Liben-Nowell 07] Liben-Nowell, D. and Kleinberg, J.: The link-prediction problem for social networks, *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 7, pp. 1019–1031 (2007)

[Nickel 11] Nickel, M., Tresp, V., and Kriegel, H.-P.: A three-way model for collective learning on multi-relational data, in *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pp. 809–816 (2011)

[Popescul 03] Popescul, A. and Ungar, L. H.: Statistical relational learning for link prediction, in *Workshop on Learning Statistical Models From Relational Data at the International Joint Conference on Artificial Intelligence*, Vol. 149, pp. 81–90 (2003)

[Rattigan 05] Rattigan, M. J. and Jensen, D.: The case for anomalous link discovery, *ACM SIGKDD Explorations Newsletter*, Vol. 7, No. 2, pp. 41–47 (2005)

*6 提案手法の AUC-ROC / 既存手法の AUC-ROC