The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

データプラットフォームにおける異分野データネットワークの成長過程に関する一考察

Network Growth Process on Cross-disciplinary Data in the Data Platform

早矢仕晃章[*1]
Teruaki Hayashi

大澤幸生[*1]
Yukio Ohsawa

[*1] 東京大学大学院　工学系研究科　システム創成学専攻
Department of Systems Innovation, School of Engineering, The University of Tokyo

Data are generated every moment, and new kinds of data appear one after another in the real world. In recent years, problem-solving by exchanging/combining data among different domains is one of the social demands. In the Web and Social Networking Services, many models to express the complex networks and their growth process have been proposed. However, the network model focusing on data from different domains in the data platform has not been discussed. In this study, we observed the process of network growth of cross-disciplinary data using the summarized information about data–data jacket–, and considered the dynamic change of the characteristics of the network in its growth process. From the characteristics of the dynamic network, we demonstrated several strategies as data platformers.

1. はじめに

実社会において，多種多様なデータが時々刻々と生成され，新たな種類のデータが続々と登場してきている．近年，データを交換可能な材として，異分野間でデータを交換・結合させることで問題解決する動きが活発になってきている．このような取引が行われるプラットフォームとして，データ流通市場，データ取引市場など，多様な形態のデータ市場及び関連サービスが萌芽してきている．データ市場では多様なステークホルダーのデータがプラットフォームに登場し，他者と交換・取引されるという点で，データをノードとしたネットワークは絶えず成長するネットワークと言える．さらに，国内外でデータプラットフォームのサービスが立ち上がりつつある中で，プラットフォーマーとしての覇権争いが行われつつある[早矢仕 19a]．しかし，どれほどの規模のデータがあればプラットフォームとして十分に機能するのか，また，どのような種類のデータをプラットフォームで取り扱うことで，自律的にコミュニティが成長するのか，ということは研究されてきていない．新たなデータがプラットフォームに参入する過程をダイナミックに捉え，モデル化することは，将来のデータのプラットフォームの活性化において重要な課題である．

Web や SNS などでは，データの複雑ネットワーク及びその成長過程を表現する様々なモデルが提案されてきた（[伏見 14][Osaka 17]など）．しかし，データ市場における異業種のデータ固有の特徴に着目したモデルは十分議論されてきたとは言い難い．そこで本研究では，データ概要情報データジャケットを用いて異種のデータのネットワークが成長する過程を観察し，成長過程におけるネットワークのダイナミックな特徴の変化について考察する．

連絡先：
早矢仕晃章，東京大学大学院工学系研究科システム創成学専攻，hayashi@sys.t.u-tokyo.ac.jp
This research was supported by JST, CREST (JPMJCR1304). 本研究にご協力いただいた共同印刷株式会社の皆様に心より感謝申し上げます．

2. 異分野データの動的ネットワーク作成

2.1 データジャケット

データジャケット（Data Jacket: DJ）は，データの概要情報を記述するためのフレームワークである[Ohsawa 13]．データに関する説明文や含まれる変数の名前（変数ラベル），保存形式，共有条件などをデータ概要情報として記述することでデータを秘匿としたまま，異なる分野のデータについて理解可能となる（図 1）．DJ はメタデータの一形態であるが，通常のメタデータと異なり，DJ は人間がデータについて理解し，議論可能とすることを目的としている．また，異なるフォーマット，異なる粒度の変数ラベルを有するデータを共通の記述ルールによって構造的にメタデータ化することで，様々な形式のデータを統一的に扱うことができる．この特徴を用いて，[早矢仕 18b][Hayashi 18]は，変数ラベル及びデータが持つコンテキスト（文脈）を介した異分野データネットワークの解析を行い，ネットワークの特徴及びデータ市場における特徴的なデータの振る舞いについて論じた．なお，変数ラベルとは，データ固有の変数を自然言語によって記述された説明文を意味する．

DJ No. XX【購買履歴データ】	
概要	東京都の◯◯スーパーマーケットで収集されている顧客の購買行動履歴。
収集方法・コスト	ポイントカードとPOSによって取得
共有条件	共有不可
データの種類	表形式、テキスト、数値
保存形式	CSV
分析・シミュレーション	時系列分析
変数ラベル	氏名、性別、顧客ID、支払金額、購入品目、日にち
分析結果	• その日の売上の計算 • 今後の売上の予測と仕入れの推定
期待される分析	顧客の購買行動とリピート率を計算し、ロイヤルカスタマーの特定が可能かもしれない。
コメント	有効なデータの組み合わせが発見されればデータの提供あるいはコラボレーションもあり得る。

図 1　DJ の記入例（[早矢仕 18a]より引用）

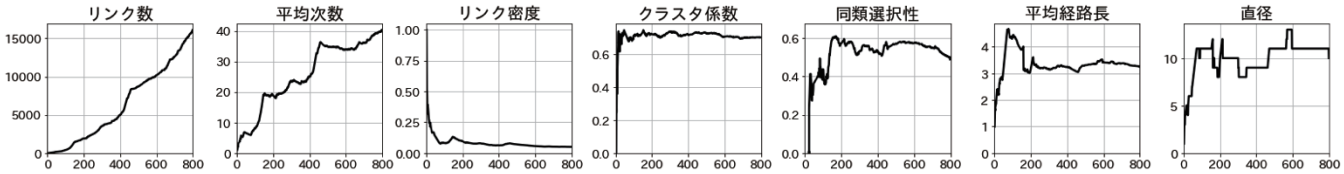

図 2　DJ（ノード数）の増加に伴う各ネットワーク特徴量の変化（横軸：ノード数）

2.2 目的と解析アプローチ

従来研究では，時間を固定した静的ネットワークを対象とした分析に留まっていた．実社会のネットワークの多くは時間変化する動的ネットワークである．そして，データ市場におけるプラットフォームでは新たなデータが参入することでデータ間の関係は時々刻々と変化するものと考えられる．ネットワークの成長過程を観察することにより，特徴的な構造変化を与えたデータの振る舞いと特徴を捉え，考察を行うことが本研究の目的である．データプラットフォームをネットワークで議論することで，ネットワークの諸問題が適用可能となり，データ同士の関係性を理解できる．変数が共通しているデータは互いに似ているという仮定によって，データの類似性を考慮したネットワークを議論できる．

本研究では，DJ ストア[1][早矢仕 16]から入手できる 1316 件の DJ のうち，ネットワークで表現した際に最大コンポーネントを構成する 798 件の DJ を対象として分析を行う．なお，ネットワークの各ノードは DJ を表し，DJ 同士が共通する変数ラベル（Variable Label: VL）を保有している場合にリンクが張られるものとした．例えば，「東京都の病院の位置情報」と「東京都に設置されている警察署の設置情報」という DJ が「緯度」と「経度」という変数ラベルを共通して有していた場合，2 つの DJ にはリンクが張られる．また，ネットワークの性質を理解する上で，ネットワークの基本統計量を用いる．本稿では，動的ネットワークの変化の特徴として，ネットワークのリンク数，各ノードの平均次数，リンク密度，クラスタ係数，同類選択性，平均経路長，ネットワーク直径の 7 項目を選んだ．

3. 結果と考察

3.1 異分野データのネットワーク

DJ の VL によるネットワークを用いることで，データのネットワークがどのように成長したのかということを時系列に従って分析することが可能となる．表 1 は 798 件の DJ の静的ネットワークの特徴量を表したものである．表にあるように，VL を介したデータのネットワークの平均次数は 40.29 と高く，リンク密度 0.051 と低い値を示している．一方，クラスタ係数は 0.702 と高い値である．リンク密度が大域的な密度を表し，クラスタ係数が局所的な密度を意味する指標であることを考慮すると，異分野データのネットワークは局所的に密なネットワークを作りやすく，大域的に密度が低い性質を持っていることが分かる．つまり，同じような変数を持つデータ同士が局所的に大きな塊を作るため，大域的に見たときに疎な構造となっているのである（図 3）．

同類選択性は隣接する 2 ノード間の次数相関を示す指標である．経験的にタンパク質などの自然界のネットワークやインターネットなどの工学系ネットワークは同類選択性が負になることが知られている．一方で，論文の共著関係や SNS などの人間関係を表すネットワークの同類選択性は正の値となりやすい．本研究の同類選択性は 0.489 と高く，同じようなデータは互いに密なネットワークを構成するものの他の領域のデータとは繋がりが疎となる特徴を持つことが分かった．また，ノード数に対して平均経路長は比較的短く，直径も小さいことから，異分野データのネットワークはスモールワールド性を有していると考えられる．

図 2 は DJ による異分野データの動的ネットワークとして捉えた際に，ノード数の増加に伴って変化するネットワークの特徴を表す指標をグラフ化したものである．リンク数は DJ の増加に伴い単調に増加しているが，あるデータの出現によってネットワーク全体のリンク数が急増する点が数箇所見られる．例えば，721 番目の「駐車場利用データ」，481 番目の「公衆トイレ情報」などはそれぞれ 160 リンク，102 リンクと，データ登場時に極めて高いリンク数を獲得している．リンク数の変化量は平均次数にも影響している．平均次数を見ると，基本的には増加しているが，リンク数の急激な増加が起こった直後に減少している部分が観察できる．これは極端に多くのリンクを獲得したデータが現れた後に見られる．極端に多くのリンクを獲得したデータが出現した直後は通常のリンク数を持つデータが乗っていくるため，ネットワーク全体の平均次数が低下するものと考えられる．

続いて，リンク密度とクラスタ係数の 2 つの指標の変化を観察すると，ノード数の増加に伴い，リンク密度は急激に低下し，一方でクラスタ係数は初期から高い値を示しており，0.70 前後の値を上下し，安定した構造を見せている．これは，データのネットワークが初期の段階から局所的に密な構造となることが分かる．また，データの追加によって密度は大きく変化することなく，局所的に密度が高い部分にてリンクを獲得していくものと考えられる．次数相関を表す同類選択性はクラスタ係数と同じく，初期から高い正の値を示しており，同じような次数のデータが互いに繋がりやすい傾向がネットワーク生成の初期から現れていることが分かる．平均経路長の変化を見ると，初期は比較的長く，その後 3.1 から 3.5 の間を推移し，安定している．ネットワーク直径もノードの増加による大きな変化はなく，10 前後を推移している．異分野データのネットワークはデータ数が少ない初期の段階から局所的に密かつ大域的に疎であり，平均次数は増加するものの他のネットワーク特徴量はデータ数 200 件ほどから急激に変化しないことが分かった．

表 1　ネットワークの特徴量

	最大コンポーネント
ノード数	798
リンク数	16076
平均次数	40.29
リンク密度	0.051
クラスタ係数	0.702
同類選択性	0.489
平均経路長	3.26
直径	10

[1] http://160.16.227.37/index/DJStore

図 3　異分野データネットワーク

3.2 データプラットフォーマーとしての戦略立案例

前節で議論したネットワークの特徴から，データプラットフォーマーとしていくつかの戦略を立案することができる．まず，自律的なコミュニティ成長においては，そのコミュニティおいて核となるデータが必要となる．すなわち，ネットワークにおいて中心性が高いデータやクラスタにおいてリンク数が高いデータが該当する．予めこのような特徴量を高くし得るデータをプラットフォームに含めておくことにより，新たなデータがプラットフォームに参入してきた際にリンクを獲得する可能性が高くなる．データ提供者にとって，自身の保有するデータがどれだけ他のデータと結合可能性が高く，周りの他のデータと比較して優位性があるのかということがデータ提供のモチベーションになるからである．

また，本稿におけるネットワークでは，同類選択性が高く，リンク密度が低い．そのため，同類選択性を低くし，リンク密度を高めるデータをプラットフォームに参入させるという戦略も考えられる．これにより，人間関係のネットワークのように局所的に強い繋がりが表れやすい「タコツボ化」が緩和し，他の領域のデータとの繋がりが生まれやすくなる．異分野データ連携を活性化させる上でプラットフォーマーは，同類選択性が高く，リンク密度が低い傾向が観察されたら，媒介中心性などを高めるデータ保有者のデータをプラットフォームに提供すると良いだろう．

4. 結論

本稿では，データ概要情報データジャケットを用いて異種のデータのネットワークが成長する過程を観察し，成長過程におけるネットワークのダイナミックな特徴の変化について考察した．本稿で扱った異分野データのネットワークは，[早矢仕 18a][早矢仕 19b]から始まった研究であり，まだまだ発展途上である．本稿のように複雑ネットワークの諸問題に落とし込むことで多様な手法が適用可能となり，データプラットフォームに関する研究のさらなる発展が期待できると考えている．

Barabási らの研究から派生し，成長するネットワークには様々なモデルや時間を考慮したテンポラルネットワークなどの動的ネットワークに関する研究が進んできた([Leskovec 05][Holme 13]など)．また，ネットワークにおける新参のノードが「下剋上」を起こして中心的な存在になることが知られている[Bianconi 01]．今後の研究では，「あるデータの登場によってどのように他のデータとの関係性が変化していくのか」という成長モデルを構築するとともに，どのような変数を含むデータを新たに設計することによって，プラットフォーム上で優位な存在となるのかということを明らかにしていきたい．また，異分野データ連携において，データの変数のみならず，文脈における共通性も重要な特徴である．データランドスケープ[早矢仕 18b]を導入したネットワークの成長過程を観察し，文脈を考慮した潜在的なデータの繋がりと動的変化を議論する必要があるだろう．また，部分的に観測されたネットワーク構造から残りの構造を推定するリンク予測問題と考えれば，潜在的なリンクを推定し，未知の VL または秘匿 VL の存在を明らかにすることができるなど，データ市場及びデータ流通プラットフォームにおいて重要な知見を得ることができると考えている．

参考文献

[Bianconi 01]　Bianconi, G., Barabási. A.L.: Bose-Einstein Condensation in Complex Networks, Physical Review Letters, Vol.86, No.24, pp.5632-5635 (2001)

[伏見 14]　伏見卓恭, 斉藤和巳, 風間一洋: PageRank に基づく動的ネットワークの構造変化抽出, 知識ベースシステム研究会, Vol.B4, No.02, pp.31-37 (2014)

[早矢仕 16]　早矢仕晃章, 大澤幸生, “Data Jacket Store: データ利活用知識構造化と検索システム,” 人工知能学会論文誌, Vol.31, No.5 (2016)

[早矢仕 18a]　早矢仕晃章, 大澤幸生: データ市場におけるデータのネットワークと関係性の分析―データの属性と繋がりからの考察―, 信学技報, 人工知能と知識処理研究会, Vol.117, No.440, pp.49-54 (2018)

[早矢仕 18b]　早矢仕晃章, 大澤幸生: データ 3.0 時代のデータランドスケープ, 2018 年度人工知能学会全国大会 (2018)

[Hayashi 18]　Hayashi, T., Ohsawa, Y.: The Difference between Variable-based and Context-based Networks of Data Using Data Jackets, 22nd International Conference on Knowledge Based and Intelligent Information and Engineering System, Vol.126, pp.1740-1747 (2018)

[早矢仕 19a]　早矢仕晃章, 小口裕, 飛沢省二, 大澤幸生: 日本・米・欧州・中国のデータ市場ビジネスの動向, 信学技報, 人工知能と知識処理研究会 (2018)

[早矢仕 19b]　早矢仕晃章, 岩永宇央, 岩佐太路, 大澤幸生: データジャケットを用いた異分野連携に資するデータの特徴とネットワーク分析, 日本知能情報ファジィ学会誌, in Press (2019)

[Holme 13]　Holme, P., Saramäki, J.: Temporal Networks, Understanding Complex Systems, Springer (2013)

[Leskovec 05]　Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. the 11th International Conference on Knowledge Discovery in Data Mining, pp.177-187 (2005)

[Ohsawa 13]　Ohsawa, Y., Kido, H., Hayashi, T., and Liu, C.: Data Jackets for Synthesizing Values in the Market of Data, 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science, Vol.22, pp.709-716 (2013)

[Osaka 17]　Osaka, K., Toriumi, F., Sugawara, T.: Effect of direct reciprocity and network structure on continuing prosperity of social networking services, Computational Social Networks, Springer, Vol.4, No.2, pp.1-20 (2017)