社員間コミュニケーションの分析を目的としたグラフ埋め込み手法に関する研究

情報数理応用研究　5219C028-9　野中賢也
指導教員　後藤正幸

A Study on Graph Embedding Method for Analysing Communications between Employees

NONAKA Kenya

1. 研究背景・目的

企業経営において，人的リソースの有効活用とパフォーマンスを最大化する職場コミュニティの醸成は大変重要な課題であり，そのため自社内の組織状態の客観的な分析・把握が求められている．このような課題に対して，E-mail の送受信データを解析することによって，社内の会話状況を把握する研究などがなされている [1]．一方で，近年，E-mail の代替ツールとして，Slack や Teams 等のビジネスチャットアプリが職場内のコミュニケーションのために使用され始めており，アプリ上に蓄積された会話履歴データの利活用により，上述の課題解決を行うことが考えられる．

チャットアプリでは，社員間の会話はチャンネルと呼ばれるプロジェクトごとに作成されたユーザーグループ上で行われる．チャンネル上での会話履歴をもとに，社員をノード，社員間の業務上の関係性の有無をエッジとして表現したグラフを構築することができる (以下，チャンネルグラフと呼ぶ)．このチャンネルグラフは，それぞれのチャンネルにおける参加者の会話傾向によって形状が異なる．例えば，一人に会話が集中するチャンネルや参加者各々が対等に会話を行うチャンネル等が存在する．そこで，部署や企業ごとのチャンネルグラフの生起パターンを分析することによって，それらの会話傾向を明らかにすることが期待できる．

しかし，各チャンネルのグラフ表現は複雑な非構造化データであり，多数のチャンネルをグラフ表現のまま分析することは困難である．したがって，各チャンネルグラフの特徴を抽出し，構造化データへと変換する必要がある．構造化データへと変換することで，従来研究されてきたクラスタリングや回帰・分類モデルといった各手法の適用が可能となる．特に本研究では，抽出された特徴ベクトルに対してクラスタリング手法を適用し，各チャンネルグラフの類型化を行う．

非構造化データの特徴抽出の方法として，ニューラルネットワークの高い表現能力を利用した埋めこみ表現モデルの研究が進められている．特に，チャンネルグラフの特徴抽出では，構造が類似したグラフを特徴空間上の近接領域に埋め込むグラフ埋め込み表現モデル [2] が有効であると考えられる．本研究では，Al-Rfou らによって提案された Deep Divergence Graph Kernel (以下，DDGK と呼ぶ)[3] に注目する．この手法は，グラフ埋めこみ表現の学習にノードやエッジのラベルを必要としない汎用性の高い手法である．DDGK では，埋め込み空間の軸に対応するグラフ (以下，ソースグラフ) を埋め込み対象のグラフ (以下，ターゲットグラフ) からのランダムサンプリングにより選定している．このような選定方法では，構成される埋め込み空間の各軸が相関を持ってしまい，また，チャンネルグラフの重要な特徴を捉えていない可能性があるため，適切とは言えない．

そこで，本研究では，適切なソースグラフを Barabási-Albert(BA) モデル [4] を用いて生成することで，チャンネルグラフの埋め込み表現を得るための特徴空間軸構成法を提案する．これにより，チャンネルグラフの重要な構造を多面的に表現できる埋め込み空間の構成が可能となる．加えて，実際の Slack 会話履歴データから提案モデルによって得られる特徴ベクトルを活用し，各チャンネルのグラフ構造を類型化し，分析することで，提案モデルの有効性を示す．

2. 準備

2.1. 対象問題

本研究で用いるデータは，チャットアプリ上に蓄積された会話履歴データである．アプリ上には，特定業務の連絡を目的としたチャンネルが多数存在し，チャンネル内で社員同士の会話が記録される．チャンネルには，2 名の参加者が直接やり取りする DM チャンネルと複数の参加者が参加するグループチャンネルが存在する．グループチャンネルでは，一つのメッセージに対して，その発信者と受信者 (当該チャンネルの参加者) が一対多で結び付いている．発信者が受信者を特定する機能 (メンション機能) を使用した場合に限り，発信者と受信者は一対一で結び付く．以上の会話履歴データから，ノードをチャンネル参加者，エッジを参加者同士のつながりとした無向重み無しグラフ $G_n(n = 1, ..., N)$ で各チャンネルを表現し，R 次元の特徴空間上の埋め込み表現 $\Psi(G_n) \in \mathbb{R}^M(n = 1, ..., N)$ を得る手法を提案する．

2.2. 従来手法

DDGK[3] は，埋め込み対象となるグラフ (ターゲットグラフ) の集合を埋め込み空間の軸に相当するグラフ (ソースグラフ) との非類似度によって，多次元特徴空間上へと埋め込む手法である．ターゲットグラフ集合を $\{G_1, \cdots, G_N\}$，ソースグラフ集合を $\{S_1, \cdots, S_M\}$ とおく．ここで，ソースグラフ集合は，ターゲットグラフからのランダムサンプリングによって事前に設定される．

ターゲットグラフ G_n の埋め込み表現 $\Psi(G_n) \in \mathbb{R}^M(n = 1, ..., N)$ の第 m 要素は，ターゲットグラフ G_n と第 m 番目のソースグラフ S_m との非類似度 $D(G_n|S_m)$ で与えられる．

$$\Psi(G_n) = [D(G_n \mid S_1), \cdots, D(G_n \mid S_M)] \quad (1)$$

非類似度 $D(G_n|S_m)$ は，ソースグラフ S_m の構造を学習したニューラルネットワーク (エンコーダ) の重み $\hat{\boldsymbol{\theta}}_m$ を用いて，ターゲットグラフ G_n の全てのノード $\boldsymbol{u}_i(i=1,\cdots,I)$ に対し，その隣接ノード $\boldsymbol{u}_j \in N(\boldsymbol{u}_i)$ を予測した際の損失として式 (2) のように定義される．

$$D(G_n \mid S_m) = -\sum_{i=1}^{I} \sum_{\boldsymbol{u}_j \in N(\boldsymbol{u}_i)} \log P(\boldsymbol{u}_j \mid \boldsymbol{u}_i, \hat{\boldsymbol{\theta}}_m) \tag{2}$$

そして，式 (1) 及び式 (2) に基づいて，ターゲットグラフ G_n の埋め込み表現が得られる．

2.3. Barabási-Albert(BA) モデル

本研究では，BA モデル [4] により生成されたグラフを DDGK のソースグラフとして用いることを考える．BA モデルは，次数分布がべき乗分布に従うネットワークを生成する代表的なモデルであり，グラフの成長と優先的選択を生成アルゴリズムに取り入れている．具体的には，ノード数 a とノード追加時のエッジ数 (初期ノード数)b を引数として以下の方法でグラフを生成する．

(I)b 個のノードを持つグラフ G を構築する．(II)G にノードを 1 個追加する (グラフの成長)．(III) 既存のノードから b 個のノードをそれぞれの次数に比例した確率で，ランダムに選択し，追加されたノードとエッジにより結びつける (優先的選択)．(IV)G のノード数が a に等しければ G を出力，そうでなければ II に戻る．

3. 提案

3.1. 埋め込み空間の軸が満たすべき要件

本研究では，社員間コミュニケーションの分析のためのチャンネルグラフ構造の類型化を行う．具体的には，チャンネルグラフの埋め込み表現にクラスタリング手法を適用する．このとき，クラスタリング手法の基礎となるデータ間の類似度には，チャンネルグラフにおける重要な構造が多面的に反映されている必要がある．データ間の類似度は，埋め込み空間の各軸に大きく依存するため，チャンネルグラフの重要な構造を多面的に評価する埋め込み空間軸が必要となる．よって，本研究目的における埋め込み空間の軸は，以下の 2 つの要件を満たす必要がある．1 点目は，チャンネルグラフにおける重要な特徴を捉えていることである．2 点目は，チャンネルグラフの多面的な形状的特徴を表現するため，埋め込み表現の各軸は互いに異なる形状を評価する，すなわち，軸同士の相関が低いことである．しかし，Al-Rfou らは，埋め込み空間の軸に対応するソースグラフをターゲットグラフからのランダムサンプリングにより設定している．すなわち，上述の要件を満たすようなソースグラフの選定は行っていない．

3.2. 要件を満たすソースグラフの生成

本研究では，上述の 2 つの要件を満たすような軸となるソースグラフをネットワーク生成モデルを用いて生成する．ネットワークの生成モデルは，生成されるネットワークの種類に応じて多数存在するが，チャンネルグラフの埋め込みには，Barabási-Albert(BA) モデル [4] によって生成されたグラフをソースグラフに用いることが適切である．その理由は，第 1 の要件であるチャンネルグラフにおける重要なネットワーク特徴の表現という観点から説明することができる．

チャンネルグラフを表現する上で，重要なネットワーク特徴を特定するため，チャンネルグラフ集合に対して，従来使用されてきたネットワーク基本特徴量 [5] のうち 13 の代表的な特徴量 (平均次数，エッジ数，最大次数，次数分散，ノード数，孤立ノード数，連結成分数，密度，次数相関，グラフ直径，平均クラスタ係数，最小次数，の 13 の特徴量．以下,「基本特徴量セット」) を計算した．これらの特徴量の共分散行列を求めて可視化したものが，図 1 である．図 1 では，各ノードが基本特徴量，エッジが特徴量間の共分散を示している．エッジの太さが各特徴量間の共分散の絶対値の大きさを示す．ただし，閾値を設け，共分散の絶対値が低いエッジは剪定した．図 1 より，チャンネルグラフを表現する重要な特徴量は，サイズに関連する特徴量及び密度に関連する特徴量の大きく 2 グループに分けられることがわかる．

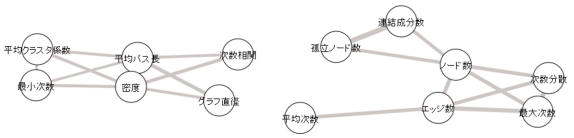


図 1: チャンネルグラフの基本特徴量間の共分散可視化

したがって，チャンネルグラフを分析するための軸となるソースグラフ集合は，サイズ及び密度が異なる様々なグラフの集合であることが望ましい．BA モデルでは，このサイズ及び密度を生成アルゴリズムの引数として扱うことができる．具体的には，2.3 で論じた BA モデルの生成アルゴリズムにおいて，ノード数 a とノード追加時のエッジ数 b という 2 つの引数の候補集合を A,B とし $|A| \times |B|$ 個の全ての組み合わせから，ソースグラフとなるグラフを生成する．このように生成されたグラフの集合は，チャンネルグラフの重要なネットワーク特徴量を捉えており，第 1 の要件を満たす．また，それぞれサイズ及び密度が異なっていおり，第 2 の要件を満たすことが期待できる．

3.3. 提案アルゴリズム

BA モデルの引数の候補集合 A,B をあらかじめ適切に設定した場合に，ターゲットグラフの埋め込み表現 $\Psi(G_n)(n = 1, ..., N)$ を学習する提案モデルのアルゴリズムを以下に示す．

1)A，B の要素すべての組み合わせに対し，BA モデルに基づき，$M = |A| \times |B|$ 個のソースグラフ集合 $\{S_1, \cdots, S_M\}$ を得る．また，ターゲットグラフのインデックスを $n = 1$ として初期化する．

2) 式 (1) 及び式 (2) から，ターゲットグラフ G_n の埋め込み表現 $\Psi(G_n) \in \mathbb{R}^M$ を得る．

3)$n = n + 1$ とする．$n \leq N$ であればステップ 2) に戻る．$n \geq N + 1$ であれば終了する．

4. 実データによる実験

4.1. データ前処理

2.1 節で論じた各チャンネルの会話履歴データからチャンネルごとの会話状況を表現するチャンネルグラフを以下の手

表 1: 各クラスの基本特徴量による特徴づけ

	クラス 1 に所属するグラフ	クラス 2 に所属するグラフ
(1) 密度	密なネットワーク	疎なネットワーク
(2) 次数相関	隣接するノード間の次数は等しい傾向	中心的なノードに次数の小さなノードが結合する傾向
(3) 最小次数	次数の小さな疎外されたノードが少ない	次数の小さな疎外されたノードが多い
(4) 平均クラスタ係数	三角的な関係が発生しやすい.	三角的な関係が発生しにくい

順により構築する.

(処理 1) チャンネル参加ユーザーが 5 名以上，週平均発言回数が 5 回以上のチャンネルを抽出する.

(処理 2) 処理 1 によって抽出されたチャンネルにおける会話履歴データをもとにチャンネルの会話状態を表す無向重み無しグラフを構築する.

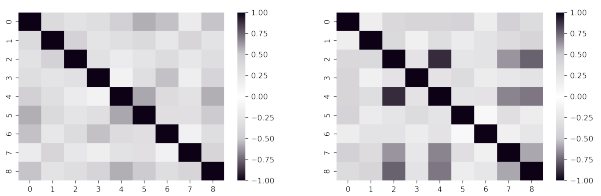
4.2. 実験条件

分析に用いた会話履歴データは，社員数 343 名，総チャンネル数 14,786 件の某企業内 Slack である．このデータに対して，4.1 節で示したチャンネルグラフの構築手法を適用し，チャンネルグラフ集合 $\{G_1, \cdots, G_N\}$ を構築した．構築されたグラフ集合は，総グラフ数 111，平均ノード数 7.52, 平均エッジ数 7.62 となった．提案手法と従来手法の詳細な実験設定は以下の通りである.

提案手法では，BA モデルの引数にあたるノード数 a 及び追加エッジ数 b の候補を $A = \{6, 9, 15\}, B = \{1, 2, 3\}$ として設定する．A, B のすべての組み合わせに対し, グラフを生成し，9 個のソースグラフを得る．従来手法では，提案手法と同数の 9 個のソースグラフを，ターゲットグラフからのサンプリングによって得る．提案手法，従来手法とも学習に使用したエンコーダの層数は $L = 4$，誤差逆伝搬法の学習率は $\alpha = 0.01$，エポック数は $e = 1000$ とした.

4.3. 相関係数行列の比較

提案手法及び従来手法で得られたチャンネルグラフの埋め込み表現に対して，その相関係数行列を計算した結果を図 2 に示す．各セルの濃度が各特徴量の相関の強さを示している．図 2(a)(提案手法) では，濃い色のセルが少なく特徴量間に強い相関が出ていないのに対して，図 2(b)(従来手法) では，特徴量 2,4,7,8 に対応するセルが比較的濃い色となっている．これは，ソースグラフをランダムサンプリングし，形状の似たグラフがサンプリングされた結果，形状の似たソースグラフに対応する軸同士に相関が発生したためである．一方で，提案手法では異なる形状を持ったソースグラフによってターゲットグラフを埋め込んだ結果，各軸の相関が低くなったと考えられる．以上より，提案手法は従来手法と比較して，相関の低い埋め込み空間の軸を構成していることがわかる.



(a) 提案手法　　(b) 従来手法

図 2: 埋め込み表現の相関係数行列のヒートマップ

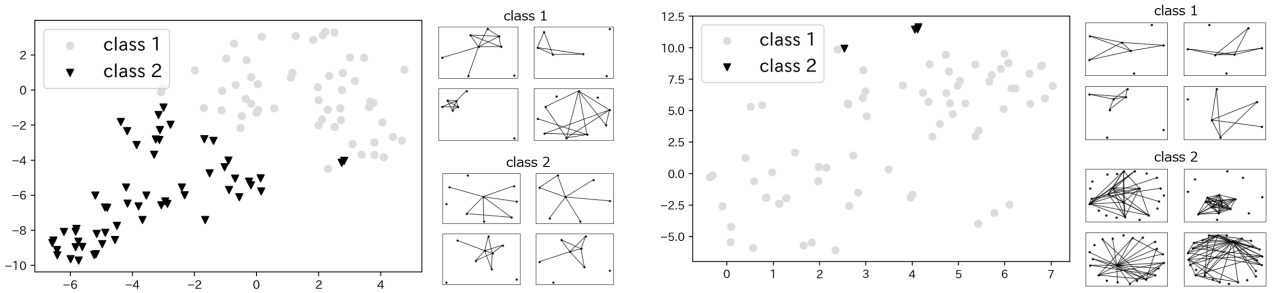
4.4. 提案手法によるチャンネルグラフの分析

チャンネルグラフの特徴をベクトル表現する上で最も単純な方法は，3.2 節で論じたネットワーク基本特徴量を計算することである．チャンネルグラフの分析において，この単純な方法と提案手法を比較する．図 3 は，提案手法及び基本特徴量セットによって計算されたチャンネルグラフの特徴量ベクトルに対し，k-means によるクラスタリングを行い，t-SNE を用いて，2 次元に写像・可視化したものである．k-means のクラス数は，k を 2 から 10 まで変化させて分析を行い，シルエット係数が最大となる $k = 2$ を選択した．また，各図の右部に可視化されているのは，各クラスの中心ベクトルから最も近い 4 つのデータ点に対応するチャンネルグラフである．図 3(b) から，ネットワーク基本特徴量に対してクラスタリングを行うことで，少数のデータが所属するクラスと多数のデータが所属するクラスに分かれることがわかる．これは，ノード数やエッジ数が大きな少数のグラフ，すなわち，外れ値的なグラフが存在するためである．したがって，グラフ基本特徴量を単純に用いてベクトル表現するだけでは，意味のあるクラスタを抽出することが難しいといえる．提案手法による埋め込み表現の図 3(a) では，図 3(b) と異なり，各クラスタに所属するデータ数に偏りは存在しない．この抽出されたクラスタが，ネットワーク基本特徴量によって特徴づけられ，解釈可能であることを以下に示す.

図 4 では，提案法の埋め込み表現から抽出されたクラス 1 及びクラス 2 に所属するチャンネルグラフについて，ネットワーク基本特徴量を計算し，クラスごとの差が大きい基本特徴量の平均を図示した．図 4 及び以下の各特徴量の性質から各クラスに所属するグラフは表 1 のように特徴づけられる.

(1) 密度：グラフ上のエッジ数を E，ノード数を V とする．グラフ上に構築することができるエッジ数 S は，V 個のノードから任意の 2 つを選ぶ全ての組み合わせ数 ${}_V C_2$ に等しい．密度は，S と E の比として計算される．密度が高ければ，グラフのサイズと比べて，多くのエッジが構築された密なネットワークとなる.

(2) 次数相関：隣接するノード同士の次数の相関により計算される．次数の高いノードに次数の低いノードが結び付くグラフであれば次数相関は低くなり，隣接するノード同士の次数が近ければ次数相関は高くなる.



(a) 提案手法による埋め込み　　(b) 基本特徴量によるベクトル表現

図 3: t-SNE による埋め込み空間の低次元可視化

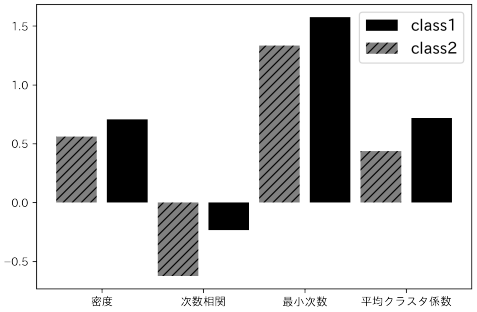


図 4: クラスごとのネットワーク基本特徴量の平均値比較

(3) 最小次数：グラフ中で最も次数の低いノードの次数である．4.1 の方法により構築したグラフは，隣接行列の平均値でエッジを剪定しており，次数が 0 のノードが発生しやすく，ほとんどのグラフで最小次数の値が 0 となる．よって，元のグラフから次数 0 のノードを除くため，最もノード数が多くなるように連結した部分グラフを抽出し，この部分グラフの最小次数を特徴量として計算した．最小次数が大きければ，次数の小さな疎外されたノードが少ないといえる．

(4) 平均クラスタ係数：ノード $\boldsymbol{v}_i$ のクラスタ係数は，ノード $\boldsymbol{v}_i$ に隣接するノード $N(\boldsymbol{v}_i)$ について，$N(\boldsymbol{v}_i)$ 間に実際に構築されたエッジ数と $N(\boldsymbol{v}_i)$ から任意の 2 つを選んで構築できるエッジ数との比で計算される．平均クラスタ係数は，グラフ上のすべてのノードに対して，上記のクラスタ係数を算出し，平均したものである．グラフ上に 3 つのノードが相互に接続された三角的な関係が発生しやすければ，平均クラスタ係数が高くなる．

会話に積極的な社員が多いチャンネルの場合，その会話履歴から構築されるグラフには多くのエッジが構築され密度が高くなり，最も会話が少ない参加者でも，ある程度の参加者と結び付いて最小次数が高くなる．また，少数の参加者に会話が集中することが少ないため，隣接するノード同士の次数すなわち次数相関が高くなる．加えて，多くのエッジが構築された結果，三角的な関係性が発生し平均クラスタ係数が高くなる傾向にあるといえる．このような密でフラットな会話が行われているチャンネルのグラフがクラス 1 として抽出されている．逆に，疎で一人集中型の会話が行われているチャンネルのグラフはクラス 2 で抽出されているといえる．

5. 考察

4.4 節における分析結果は，各チャンネルのチームマネージャーに対する現状の可視化や意思決定支援に有用であると考えられる．例えば，自身がマネジメントするチームのチャンネルグラフのクラスが一人集中型である場合，よりフラットな関係性を目指し，メンバー間が気軽に意見の出せる雰囲気作りのための施策を打つ動機となり得る．また，同一チャンネルの施策実施以降の会話履歴データから構築されたチャンネルグラフを同一空間内に埋め込むことで，施策の結果，実際にフラットな関係性が構築できたかを確認できる．

6. まとめと今後の課題

本研究では，グラフ生成アルゴリズムによって埋め込み空間の軸に相当するグラフを生成し，Deep Divergence Graph Kernel のソースグラフを与えるグラフ埋め込み手法を提案した．さらに，実データに適用し，その分析によって提案手法の有効性を検証した．今後の課題としては，部署・プロジェクト名等のチャンネル属性情報を活用し，それらの属性情報と埋めこみ表現のクラスタリングから得られるグラフ形状に関する情報を統合的に分析することなどが挙げられる．

参考文献

[1] H. Yang, J. Luo, Y. Liu, M. Yin and D. Cao, “Discovering Important Nodes through Comprehensive Assessment Theory on Enron Email Database, *2010 3rd International Comference on Biomedical Enginneering and Informatics,* Vol. 7, pp.3041-3045, 2010.

[2] H. Cai, V.W. Zheng and K.C. Chang, “A Comprehensive Survey of Grpah Embedding: Problems, Techniques, and Applications,” *IEEE Transactions on Knowledge and Data Engineering,* Vol.30, No.9, pp. 1616-1637, 2018.

[3] R. Al-Rfou, D. Zelle, and B. Perozzi, “DDGK: Learning Graph Representaions for Deep Divergence Graph Kernel,” *Proceedings of the 2019 World Wide Web Conference on World Wide Web,* pp. 37-48, 2019.

[4] A.L. Barabasi and R. Albert, “Emergence of Scaling in Random Networks,” *Science*, Vol.286, pp. 509-512, 1999.

[5] L. da F. Costa, F. Rodrigues, G. Travieso, and P. Boas, “Characterization of Complex Networks: A Survey of Measurements,” *Advances in Physics*, Vol.56, pp. 167-242, 2007.