ビジネスチャット上の会話内容に着目した社員間の関係性の可視化分析

情報数理応用研究　5220C011-3　川上達也
指導教員　後藤正幸

Visualization Analysis of Relationships between Employees Focusing on Content of Communication on Business Chat

KAWAKAMI Tatsuya

1. 研究背景・目的

ノードを各社員として社員間の関係をネットワーク構造で表現する企業ネットワーク分析は，密に繋がりを持つコミュニティや影響のある社員の発見を可能とする．しかし，従来のネットワーク分析 [1] では， E-mail 等のコミュニケーションツールにおけるメッセージの送受信回数に基づいて構築したネットワークを対象としており，具体的な会話内容に踏み込んだ議論はされていなかった．そこで，本研究ではビジネスチャットにおける会話内容を考慮したネットワーク可視化分析手法を提案する．

ここで，エッジを表す社員間の繋がりの定義によって，得られるネットワーク構造が変化することに着目する．例えば，ビジネスチャットでは参加ユーザから構成されるグループが存在しており，各グループのメッセージの送受信回数によりエッジの有無を定義すると，グループごとに大きく異なるネットワークが得られる．このように，本研究で対象とする社員ネットワークでは各ノードが表す社員は同じであるが，発言内容やトピック，所属グループなどのエッジの定義に用いる会話の条件（コミュニケーション条件）によってノード間の関係が変化する．これは，従来のネットワーク分析で対象とする SNS 上での友人関係などのネットワークとは異なる点である．また，コミュニケーション条件の違いに伴うネットワーク構造の変化や，全体的なコミュニケーション構造と大きく異なる構造を持つコミュニケーション条件を解釈しやすい形で可視化することは有意義であるといえる．

しかし，ネットワークの可視化による社員の関係性分析に際して，各ノードの二次元空間上の位置は Kamada-Kawai Algorithm(以下，KK Algorithm)[2] などで数理的に決定される．そのため，独立に作成された複数のネットワーク間で対応するノードの位置が異なってしまい，ネットワーク構造の比較は困難となる．そこで，本研究ではコミュニケーション条件の違いによる社員ネットワーク構造の変化を定量的に分析し可視化を行う，ビジネスチャットのためのネットワーク分析手法を提案する．このとき，KK Algorithm のエネルギー関数をネットワーク間の乖離度指標として用いることで，全体的なコミュニケーション構造から乖離した，特徴的なネットワーク構造を持つコミュニケーション条件を発見することが可能となる．最後に実際の会話履歴データを用いた分析により，本手法を用いた社員間の関係性の分析が有効であることを示す．

2. Kamada-Kawai Algorithm

Kamada-Kawai Algorithm[2] は，ノード間の距離に基づいた引力と斥力を仮定することで各ノードの最適な位置を決定するネットワーク可視化手法である．具体的には，全ノード間に仮定したばねのエネルギーの総和が最小となるようにノードを配置することで，ノード間の最短経路長と座標のユークリッド距離が一致するようにノードの位置を決定する．

ここで，連結グラフ $\boldsymbol{G}$ における N 個のノード集合を $\boldsymbol{V} = \{v_i\}_{i=1}^{N}$，各ノード v_i の位置を $\boldsymbol{p}_i = [p_{i1}, p_{i2}]$，$N$ 個のノードの位置集合を $\boldsymbol{P} = \{\boldsymbol{p}_i\}_{i=1}^{N}$ と表す．このとき，全ばねにおけるエネルギーの総和は，以下の式 (1) で表される．

$$E(\boldsymbol{P}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{1}{2}k_{ij}\left(\left(p_{i1}-p_{j1}\right)^2 + \left(p_{i2}-p_{j2}\right)^2 + l_{ij}^2 - 2l_{ij}\sqrt{\left(p_{i1}-p_{j1}\right)^2 + \left(p_{i2}-p_{j2}\right)^2}\right) \quad (1)$$

式 (1) における k_{ij},l_{ij} はそれぞれノード v_i とノード v_j 間のばね定数，ばねの理想距離を表す定数であり，ノード v_i とノード v_j の最短経路長 d_{ij} を用いて以下の式 (2)，(3) により求められる．

$$k_{ij} = \frac{K}{d_{ij}^2} \quad (2)$$

$$l_{ij} = L \times d_{ij} \quad (3)$$

ただし，K は正定数，$L = L_0/\underset{i<j}{max}\ d_{ij}$ は定数 L_0 を用いて計算される値である．これにより，グラフ理論に基づくネットワーク構造とノード間のユークリッド距離が対応したノードの位置を得ることができる．また，式 (1) における総エネルギーの最小化では，エネルギーの減少幅が最大となるノード v_l を選択し，Newton-Raphson 法 [3] による更新を繰り返すことでノードの位置の最適化を行う．

3. 提案手法

3.1. 提案手法の概要

本研究では，社員間のコミュニケーション構造の分析を目的として，重みなし無向グラフである社員ネットワークを対象としたネットワーク可視化分析手法の提案を行う．具体的には，全会話履歴データを基に組織の全体的なコミュニケーション構造を表すネットワークである「ベースネットワーク」を作成し，KK Algorithm により最適なノードの位置を数理的に決定する．その後，ノードの位置を固定した状態で，発言内容やトピック，所属グループなどの会話をカウントする条件を表す「コミュニケーション条件」に基づいてネットワークを作成する．このネットワークは各コミュニケーション条件と一体一対応しており，本研究では「条件付きネットワーク」と定義する．これにより，コミュニケーション条件の違いに伴うネットワーク構造の変化をエッジの増減で捉

えることが可能となる．さらに，KK Algorithm のエネルギー関数を各条件付きネットワークのベースネットワークからの乖離度として用いることにより，全体的なコミュニケーション構造から大きく構造が変化するコミュニケーション条件を発見する．

3.2. 変数定義とノードの二次元空間への写像

社員数を N，各社員 i に対応するノードを $v_i \in \boldsymbol{V}$ と設定する．ただし，各ネットワーク間におけるエッジ集合の変化に伴う孤立ノードの発生を考慮し，ノード集合 $\boldsymbol{V}$ に孤立ノードが含むことを許容する．また，ベースネットワークのエッジ集合を $\boldsymbol{E}_{base}$，ベースネットワークを $\boldsymbol{G}_{base} = (\boldsymbol{V}, \boldsymbol{E}_{base})$ と表記する．加えて，考慮可能なコミュニケーション条件の数を M と置く．このとき，発言単語やトピックをコミュニケーション条件 m として選択すると，エッジ集合が $\boldsymbol{E}_m$ へと変化し，対応する条件付きネットワーク $\boldsymbol{G}_m = (\boldsymbol{V}, \boldsymbol{E}_m)$ が得られる．さらに，全ネットワーク間で共通の各ノードの位置集合を $\boldsymbol{P} = \{\boldsymbol{p}_i\}_{i=1}^{N}$，ノード v_i の二次元空間上の位置を $\boldsymbol{p}_i = [p_{i1}, p_{i2}]$ と設定すると，ベースネットワーク $\boldsymbol{G}_{base}$ は $(\boldsymbol{G}_{base}, \boldsymbol{P})$，条件付きネットワーク $\boldsymbol{G}_m$ は $(\boldsymbol{G}_m, \boldsymbol{P})$ を用いて二次元空間上に表示することが可能となる．

3.3. 提案手法の手順

具体的な提案手法の分析アルゴリズムを以下に示す．

Step 1: ベースネットワークの作成

会話履歴データ上の全発言における，社員 i から社員 j へのメッセージ数 s_{ij} と社員 j から社員 i へのメッセージ数 s_{ji} の合計を，社員 i と社員 j 間のコミュニケーション強度を表す特徴量 S_{ij} とすることで特徴量行列 $\boldsymbol{S}$ を得る．

その後，k-NN グラフ法 [4] によりベースネットワークの構築を行う．k-NN グラフ法では，特徴量 S_{ij} が上位 k 個に含まれるようなノード v_i とノード v_j の間にエッジ (v_i, v_j) を引くことで，k 個のエッジ集合 $\boldsymbol{E}_{base}$ を得る．これにより，ベースネットワーク $\boldsymbol{G}_{base} = (\boldsymbol{V}, \boldsymbol{E}_{base})$ が定義される．

Step 2: KK Algorithm に基づくノードの位置決定

Step 1 で得られたベースネットワーク $\boldsymbol{G}_{base}$ のノード集合 $\boldsymbol{V}_{base}$ とエッジ集合 $\boldsymbol{E}_{base}$ を用いて，KK Algorithm により各ノードの位置を決定する．KK Algorithm では連結グラフ $\boldsymbol{G}$ が与えられたときに，式 (2), (3) により算出したばねの理想距離 l_{ij}，ばね定数 k_{ij} を定数として，式 (1) のエネルギーの最小化によりノードの位置集合を決定する．提案手法では，KK Algorithm を用いてベースネットワークの連結成分に含まれるノード集合 $\boldsymbol{V}_{connect}$ の位置 $\boldsymbol{P}_{connect} \subset \boldsymbol{P}$ の最適化を行う．その後，残りの孤立ノード集合 $\boldsymbol{V}_{isolate}$ の位置 $\boldsymbol{P}_{isolate} \subset \boldsymbol{P}$ を決定する．これは，連結グラフを対象とした KK Algorithm では，孤立ノードを含む非連結成分中のノードの位置を決定できないためである．ただし，$\boldsymbol{P} = \boldsymbol{P}_{connect} \cup \boldsymbol{P}_{isolate}$，$\boldsymbol{V} = \boldsymbol{V}_{connect} \cup \boldsymbol{V}_{isolate}$ を満たすとする．

Step 3: 条件付きネットワークの作成

Step 2 で得られたベースネットワークのノードの位置 $\boldsymbol{P}$ を用いて，選択したコミュニケーション条件ごとに条件付きネットワークを構築し，その構造を可視化する．具体的には，任意のコミュニケーション条件 m に関連するメッセージを抽出することで社員のコミュニケーション強度を表す特徴行列 $\boldsymbol{S}'$ を再計算する．その後，特徴行列 $\boldsymbol{S}'$ を用いて k-NN グラフ法により，条件付きネットワークのエッジ集合 $\boldsymbol{E}_m$ を得る．これにより，条件付きネットワーク $\boldsymbol{G}_m = (\boldsymbol{V}, \boldsymbol{E}_m)$ が定義される．このとき，各ノードの位置及び，エッジ数 $|\boldsymbol{E}|$ をベースネットワークに固定することで，エッジの変化でネットワーク構造の違いの分析が可能となる．

Step 4: エネルギーによるベースネットワークからの乖離度算出

Step3 で得られた条件付きネットワーク $\boldsymbol{G}_m$ に関して，KK Algorithm を用いた評価を行う．条件付きネットワークでは，各ノードの位置 $\boldsymbol{P}$ が固定され，エッジ集合 $\boldsymbol{E}_m$ がベースネットワークから変化している．そのため，ノード間 (v_i, v_j) の最短経路長がベースネットワークから変化しており，それに伴ってばね定数 k_{ij}，ばねの理想距離 l_{ij} が変化している．よって，ノードの位置集合 $\boldsymbol{P}$ を定数，ばね定数 k_{ij}，ばねの理想距離 l_{ij} を引数として，条件付きネットワーク $\boldsymbol{G}_m$ の KK Algorithm に基づくエネルギーを求めることができる．このときのエネルギー関数を式 (4) に示す．

$$E(k_{ij}, l_{ij}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{2} k_{ij} (|\boldsymbol{p}_i - \boldsymbol{p}_j| - l_{ij})^2 \quad (4)$$

式 (4) のエネルギーが大きいほどベースネットワークから構造が変化している．そのため，大きなエネルギーを持つコミュニケーション条件に着目することで，特異なコミュニケーション構造を持つ業務の把握が可能となる．

4. 人工データを用いた実験

本節では複数の異なる性質を持つネットワークを人工的に作成することで，設定した真のネットワーク構造を提案手法が抽出可能か否かについて定量的な評価を行う．

4.1. 人工データの生成過程

各社員が所属部署に応じたコミュニケーションを取ると仮定して，以下の過程により人工データの生成を行った．

Step 1: 各ノードの特徴量の生成

各ノード v_i について K 次元の特徴量 $\boldsymbol{w}_i$ を生成する．ここで，各ノード v_i は 1 つの部署に所属しており，特徴量 $\boldsymbol{w}_i$ を K_{dep} 次元の部署ごとの特徴の結合で定義する．まず，ノード v_i の特徴量 $\boldsymbol{w}_i$ の各要素の生成方法を定める割り当て $\boldsymbol{a}_i$ を決定する．$\boldsymbol{a}_i$ の第 k 要素 a_{ik} は，所属部署に対応する k では $\{P(0) = 1 - p_{in}, P(1) = p_{in}\}$，他の k に対しては $\{P(0) = 1 - p_{other}, P(2) = p_{other}\}$ の確率で，$\{0, 1, 2\}$ から決定される．各ノード v_i の k 番目の特徴量 w_{ik} は割り当て a_{ik} を用いて以下の式 (5) により生成される．

$$w_{ik} \sim \begin{cases} \mathrm{Exp}(\lambda) & (a_{ik} = 0) \\ \mathcal{N}(\mu_{in}, \sigma) & (a_{ik} = 1) \\ \mathcal{N}(\mu_{other}, \sigma) & (a_{ik} = 2) \end{cases} \quad (5)$$

Step 2 : エッジの有無の決定

ノードペア (v_i, v_j) について，特徴量の内積 $\boldsymbol{w}_i \boldsymbol{w}_j^{\top} = W_{ij}$ が上位 $|\boldsymbol{E}|$ 件に含まれる場合，エッジを持つとする．

4.2. 生成条件

本実験では，以下の 4 パターンのネットワークを仮定して生成を行った．

- ラベル A：部署内の会話が中心のネットワーク群
- ラベル B：部署間の会話が中心のネットワーク群
- ラベル C：2 部署から構成されるネットワーク群
- ラベル D：1 部署から構成されるネットワーク群

ただし，ラベル間で共通のパラメータとして全ノード数を 160，全部署数を 4，部署ごとの所属人数を 40，$K_{dep} = 4, \mu_{in} = 10, \sigma = 0.5, \lambda = 1$ と設定し，各ラベルごとに 200 個のネットワークを生成した．また，各ラベルにおいて生成に用いたその他のパラメータを以下の表 1 に示す．

表 1: 各ラベルにおけるパラメータ

ラベル	部署数	$\|\boldsymbol{E}\|$	p_{in}	p_{other}	μ_{other}
A	4	300	0.30	0	10
B	4	300	0.10	0.40	5
C	2	300	0.50	0.01	10
D	1	200	0.10	0	10

4.3. 実験条件

提案手法では，ベースネットワークのエッジ数を 1,000 と設定し，各ネットワークにおけるエネルギーとエッジを持つノード数を元に，Gaussian Mixture Model (GMM)[3] によるクラスタリングを行った．また，比較手法としてネットワーク埋め込み手法である Netsimile[5] を用いて，得られた 35 次元の特徴量に対して GMM を適用した．

評価については，ラベル A〜ラベル D の 4 種類を正解ラベルとして Adjusted Rand Index (ARI)[6] を用いた．実験では，ネットワークの生成および各手法を用いたクラスタリングを 100 回行い，得られた ARI の平均により評価を行った．

4.4. 実験結果

各手法における ARI の値を表 2 に示す．

表 2: 人工データに対するクラスタリングの評価

	Netsimile	提案手法
ARI	0.8800	**0.9299***

$^*p < 0.05$

表 2 より，提案手法から得られるネットワークの埋め込み表現で，より精度よくクラスタリングできている．このことから，Netsimile よりも少ない次元数で優れた埋め込み表現を得られているとわかる．

5. 実データを用いた分析

本章では，実際のビジネスチャット上の会話履歴に着目し，提案手法により会話内容から得られたネットワークの可視化を行うことでその有効性を示す．

5.1. 分析対象データ

本分析では，Laboratik 株式会社提供のある実企業における Slack 上の会話履歴データを用いる．データ収集期間は 2019 年 5 月 7 日から 9 月 30 日である．また，総社員数は 325 人，グループ数は 14,742 (ダイレクトメッセージを含む)，総発言数は 1,159,335，部署数は 9 である．

5.2. 分析条件

本研究では，発言中の単語をコミュニケーション条件とし，頻出上位 500 単語を分析対象のデータとした．そのため，作成される条件付きネットワークの数 M は 500 である．また，ベースネットワークを含む各ネットワークのエッジ数 $|\boldsymbol{E}|$ を 500 と設定した．加えて，条件付きネットワーク $\boldsymbol{G}_m$ が非連結グラフである場合は，孤立ノードやサイズが 3 以下の連結グラフの除外後に連結グラフへと変換した．KK Algorithm については，$K = 1, L_0 = 1$ と設定し，ばね定数 k_{ij} とばねの理想距離 l_{ij} を算出した．

5.3. 分析結果

5.3.1. ベースネットワークの分析

全会話履歴から作成されたベースネットワークを図 1 に示す．ここで図中で同じ色のノードは同じ部署であることを表す．また，ベースネットワークのノード数は 232，エネルギーは 44.84 になっている．図 1 からベースネットワークでは同一部署間のエッジが多く，部署内でのコミュニケーションが中心であるといえる．従って，ベースネットワークから構造が大きく変化したネットワークでは，異なる部署間のコミュニケーションが増えていると考えられる．

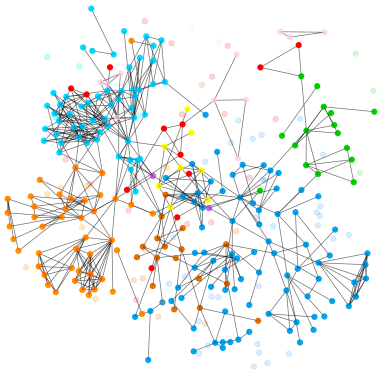

図 1: KK Algorithm を適用し得られたベースネットワーク

5.3.2. ネットワークの詳細分析

本項では，KK Algorithm のエネルギー関数を用いた，各条件付きネットワークのベースネットワークからの乖離度に関する分析を行う．各条件付きネットワークにおけるエッジを持つノード数とエネルギーの値の関係を図 2 に示す．図 2 において，青色の点はコミュニケーション条件として選択された単語 m に対応する条件付きネットワークを表し，赤色の点はベースネットワークを表している．

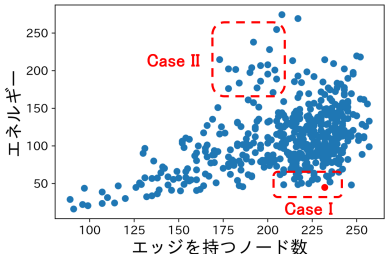

図 2: 各条件付きネットワークにおけるエネルギーとエッジを持つノード数の関係性

ここでは，図 2 に示す 2 つのケースについて実際のネットワークを表示して分析を行う．また，それぞれのケースに含まれる単語集合を以下の表 3 に示す．

- *Case I*: エネルギーが小さいが，ノード数が多いネットワーク集合
- *Case II*: エネルギーが非常に高いネットワーク集合

表 3: 各ケースにおける単語リスト

Case	単語リスト
Case I	行く, 来る, 戻る，言う, 伝える, 入る
Case II	ランク, 配信, 取材, 広島, 社員名

これらの 2 つのケースに含まれる単語集合の中から,「言う」(*Case I*),「ランク」(*Case II*) を選択し，ネットワークの可視化を行った．まず *Case I*:「言う」(図 3) に着目すると，ネットワークのエッジを持つノード数は 226，エネルギーは 60.75 である．このネットワークはベースネットワークと非常に類似しており，また表 3 の単語リストについても全社的に使用されている単語が多く，分析により得られる知見は少ないと考えられる．

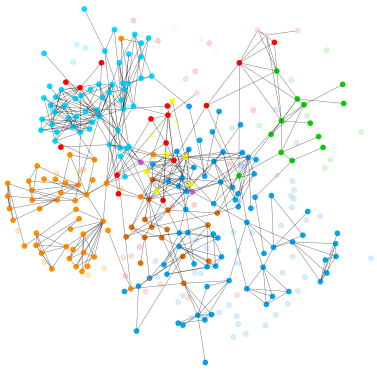

図 3: *Case I*:「言う」に基づくネットワーク

次に，*Case II*:「ランク」(図 4) に着目すると，ネットワークのエッジを持つノード数は 196，エネルギーは 205.30 である．このネットワークでは次数の高いノードが複数存在しており，それらのノードは離れたノードに多数のエッジを持っている．表 3 の単語リストに着目すると，特定の名前や地名に会話が集中しており，固有名詞が多く含まれていることがわかる．

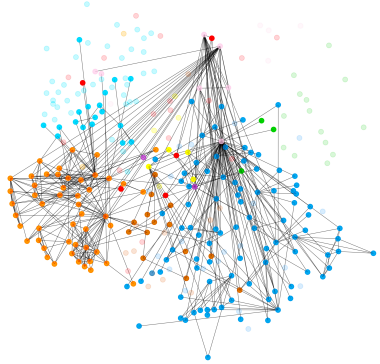

図 4: *Case II*:「ランク」に基づくネットワーク

これらの結果から，エネルギーの低いネットワークでは大半のエッジの距離が短く，エネルギーの高いネットワークでは他部署の多数の社員と横断的にコミュニケーションを取る社員が複数存在しており，ネットワーク中心性の高い社員を特定することが可能である．

6. 考察

本研究では全会話履歴からベースネットワークを作成することで，全体的なコミュニケーション構造からの乖離を評価する手法を提案した．この各ノードの位置決定に用いるネットワークを分析観点に応じて変更することで，更なる知見がを得ることができる．例えば，特定のコミュニケーション条件から得られるネットワークを用いて KK Algorithm によるノードの位置決定を行うことで，そのコミュニケーション条件と類似したコミュニケーション条件の特定が可能となる．

また，実データ分析ではベースネットワークのエッジ数を 1,000 と設定したが，少なすぎるエッジ数では孤立ノードが増加して全体的なコミュニケーション構造を捉えることができない．一方，多すぎるエッジ数ではコミュニケーションの強弱を考慮していないネットワーク構造となるため，適切なエッジ数の設定が提案手法において肝要であるといえる．

7. まとめと今後の展望

本研究では，コミュニケーション条件の違いによる社員ネットワーク構造の変化を定量的に分析し，可視化を行う手法を提案した．結果として，ビジネスチャットにおける社員間の関係性分析がより細かい粒度で行うことが可能になり，本研究の意義を結論付けることができた．今後の課題としては，ベースネットワークから大きく乖離したネットワークの自動抽出が挙げられる．

8. 謝辞

本研究にあたり，貴重なデータの提供と様々なアドバイスを頂いた Laboratik 株式会社の皆様に深く感謝致します．

参考文献

[1] Huijie Yang, Junyong Luo, Yan Liu, Meijuan Yin, and Ding Cao. Discovering important nodes through comprehensive assessment theory on enron email database. In *International Conference on Biomedical Engineering and Informatics*, Vol. 7, pp. 3041–3045, 2010.

[2] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing letters*, Vol. 31, No. 1, pp. 7–15, 1989.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] Ferreira Leonardo and Zhao Liang. A time series clustering technique based on community detection in networks. *Procedia Computer Science*, Vol. 53, pp. 183–190, 2015.

[5] Michele Berlingerio, Danai Koutra, Tina Eliassi-Rad, and Christos Faloutsos. Network similarity via multiple social theories. In *ASONAM*, p. 1439–1440, 2013.

[6] Douglas Steinley. Properties of the hubert-arable adjusted rand index. *Psychological methods*, Vol. 9, No. 3, p. 386, 2004.