千葉大学大学院融合理工学府
修士論文

Twitter分析によるSNS上の情報拡散原理と拡散過程のモデル化

平成 31 年 3 月　提出

地球環境科学専攻　都市環境システムコース

南川　雅人

Abstract

Due to the spread and development of the Internet and smart phones, anyone can easily send and receive information on WEB. From there, the development of SNS (Social Networking Service) which creates a community where people and people share information like a real society is remarkable. Along with that, there are positive aspects in which SNS is utilized as a place for corporate activities and election activities, and negative aspects in which the spread of hoax and criticism of individuals and companies (flames) are carried out. So it has great influence on our daily life.
From these facts, it is required to clarify the characteristics of SNS and clarify information diffusion process and its behavioral principle.
In this research, target SNS was set to Twitter and tweets were collected by keyword search using Twitter API.
Based on the analysis result of the tweet acquired by the keyword, (1) the tweet can be classified by the ratio of the number of retweets among all the tweets acquired the keyword (2) the number of tweets declined after reaching the peak, regularly according to day and night The time fluctuation affects the number of tweets (3) common to tweets collected with different keywords. The number of retweets obey long tail distribution (4). We identified the metadata correlated with the number of retweets. Explain that the information diffusion process can be visualized simply by reproducing actually acquired data by simulation model.

概要

インターネットやスマートフォンの普及，発達により誰でも気軽に WEB 上へ情報を発信，獲得できる社会となっている．そこから，現実社会のように人と人が情報を共有しあうコミュニティを作り出す SNS(Social Networking Service)の発展も著しい．それに伴い，SNS が企業活動や選挙活動の場として利活用される正の側面と，デマの拡散や個人，企業への非難（炎上）が行われる負の側面があり，私たちの意思決定や日常に大きな影響を与えている．これらの事実から SNS の持つ特性を明らかにし，情報拡散過程とその行動原理を明らかにすることが求められている．

本研究では対象 SNS を Twitter とし，Twitter API を用いてキーワード検索によるツイートの収集を行った．

キーワードで取得したツイートの分析結果から，（1）ツイートにはキーワード取得した全ツイートの内のリツイート数の割合による分類ができること（2）ツイート数はピークを迎えた後減衰し，昼夜に従う規則的な時間変動がツイート数に影響していること（3）異なるキーワードで収集したツイートに共通して．リツイート数は裾の長い分布に従うということ（4）リツイート数と相関関係のあるメタデータの特定を明らかにした．

実際に取得したデータをシミュレーションモデルにより再現することで簡易的に情報拡散過程を可視化することができることを説明する．

目次

第 1 章　序論

本章では，昨今の SNS（Social Networking Service）の利用実態から見る今後の SNS の活用方法や情報拡散過程に関する既存研究の変遷を通じて，本研究の背景と目的を述べる．

1.1　研究背景

スマートフォンの普及や移動通信システムの発達により現代では人々がより気軽に情報を受発信できる環境にある．特に SNS と呼ばれる会員制のウェブコミュニティではユーザー間で活発に情報のやりとりがタイムリーに行われている．SNS が発達する以前はマスメディアや口コミが情報源であったが, SNS 上の情報は以前とは比べようもないほど広範囲, 高速に情報を受け取ることが出来るとあって人気を博している．SNS 上で話題になりヒット作が生まれるといった例も見られる．このような SNS の拡散力,即時性を活用して企業のマーケティング活動の場，災害時には公式,非公式問わず有益な情報が得られるプラットフォームとなってきている．東日本大震災時には救助要請やデマ情報の拡散など我々の行動に Twitter 上で扱われる情報が影響した．[1]

このように私たちの生活に影響力を持ち始めている SNS 上の情報拡散の要因を明らかにすることが求められている．研究対象として大きく 3 つに分けられる．“どのように拡散が進むのか”という拡散のダイナミクスについての研究，“誰が拡散の重要人物であるか”というネットワーク上のインフルエンサー特定の研究，“どのような内容が拡散されるのか”という情報の価値を焦点に研究が進められている．

1.2　研究目的

背景を踏まえ，Twitter 上に投稿されている実データを取得し，どのように拡散現象が起きているか時間変化とともに図示することで情報拡散のダイナミクスを可視化すること，得られたデータを元にリツイート行動モデルを構築し，リツイート数の時間変化や分布特性の再現を目指す．

第 2 章　関連事項

この章では本研究で対象とした SNS である Twitter について説明し，ツイートデータ収集のために用いた Twitter API にも簡単に触れる．
また，SNS を対象とした関連研究について紹介する．

2.1　Twitter について

インターネット上でコミュニティネットワークを形成するサービスのことを SNS（Social Networking Service）という．Twitter は日本語 140 文字以内でつぶやき（ツイート）を投稿する SNS の 1 つである．他ユーザーをフォローすることで，そのユーザーのツイートを自身のタイムライン上で閲覧することが可能になる．自分をフォローしている他ユーザーをフォロワーと呼ぶ. 投稿に対する引用投稿（リツイート）機能や投稿記事に賛同を示す「いいね」機能などが主な機能である．Twitter では短文を気軽に投稿できることからリアルタイムな情報が多いという点が他の SNS と大きく異なる点である．現在アクティブユーザー数は日本で 4500 万人超，全世界で 3 億 2000 万人を超える巨大サービスである．従来はコミュニティ内のコミュニケーションの場として使われていたが，国内では 2011 年 3 月に発生した東日本大震災をきっかけに大規模災害時の被害状況や避難情報などの情報伝達の場としての側面を持ち始めた． Twitter の持つ情報の即時性や広範囲に情報を伝搬させることから，企業のマーケティングや米大統領選挙のようなプロモーション活動などにも活用事例がある. 一方，不安を煽るような間違ったデマ情報が拡散してしまうという危険な側面も持ち合わせている．[2]

2.2　Twitter API について

本研究で Twitter 上のツイート，付随してフォロワー数などのメタデータを取得する際に用いた Twitter API について説明する．

Twitter API とは，Twitter データにプログラミングレベルでアクセスする際に必要となるインターフェイスである. 利用にはアプリケーション登録を必要とする. 種類は大きく分けて REST API と Streaming　API の 2 つがある．

2.2.1 REST API

REST API とは REST(Representational Slate Transfer)の原則に沿って設計された Web API の 1 種である．REST の原則とは

1. リソースを一意な URI で表現すること
2. HTTP 技術をベースとし，HTTP メソッドで操作方法を表現すること
3. 処理結果をコードで表現すること

の 3 つである．ユーザーが必要とするリソースを URI で表現し，そのリソースに対しクライアントは必要な操作を GET や POST といった HTTP メソッドで示すことで処理を実行し，サーバーから処理結果を受け取る．POST でツイートの作成，削除，リツイートなど主に情報の更新を行い，GET でツイート，いいね数などの情報の取得を行う．Twitter REST API には一度に取得できるツイート数は 100 件，アクセス制限が 15 分間に 180 回までという制限がある．

2.2.2 Streaming API

Streaming API とは Long polling と呼ばれる技術を利用して疑似的にサーバーからクライアントへの情報プッシュを行う方式である． Streaming API を用いることでツイートをリアルタイムに取得することが出来る．無料，有料サービスがあり，無料版では全ツイートの 1 パーセント程度にサンプリングされたツイートデータを取得する．ただし，無料版は 2018 年 8 月にサービス停止しており現在は企業，研究機関を対象とした有料版のみの展開となっている．

2.3 ツイートデータについて

本研究では REST API を用いたキーワード検索によるツイートデータの収集を行った．収集期間は 2018 年 6 月から 2019 年 1 月までとし，ツイートデータは適宜本論文に追加した．キーワード検索では指定したキーワードを本文に含むツイートが収集できる．
得られたツイートデータのフォーマットイメージ（一部抜粋）を図に示す．また，各データが示すものを簡単に解説する．

status_code は API 呼び出しの正常/異常判定を行うものであり正常終了でコード 200 を返す. X-Rate-Limit-Remaining はアクセス可能回数を示す. この回数が 0 になると API 使用制限になり，X-Rate-Limit-Reset（アクセスリセット時間）が 0 になるまで使用できない．Text はツイート本文で id はツイート ID をあらわし各ツイートに固有のものである．Screen_name でユーザー表示名, Followers_count でユーザーのフォロワー数, Created_at でツイート投稿日を示す. 取得したツイートがリツイートの場合，オリジナルのツイート情報を Retweeted_status で得ることが出来る．

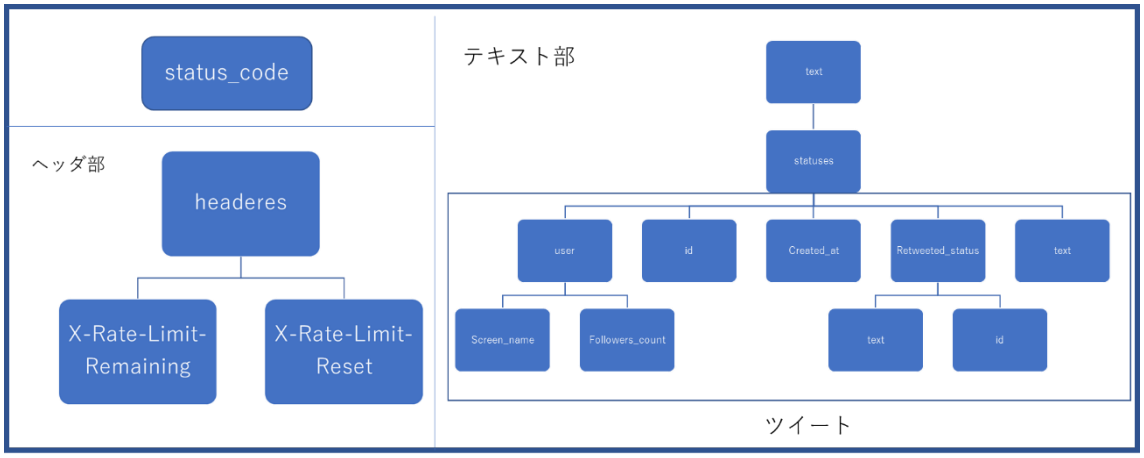

図 2.1：ツイートデータのフォーマットイメージ

2.4 関連研究

前章で述べた通り，インターネットやスマートフォンの普及，発達によりオンライン上での情報の拡散速度は増し，範囲も広範囲となっている．それ故，SNS のメディアとしての重要性が高まっており SNS を対象とした情報拡散の解析は重要な問題となっている．SNS 上の情報拡散における既存研究は大きく以下の 3 つに分類される．

1 つ目は情報の拡散と減衰のパターン解析である．ネットワーク上のあるユーザーから発信された情報が他のユーザーを介して伝播された後，やがて情報拡散は減衰し収束する．この一連のダイナミクスを解明するものがある．Youtube を対象とした先行研究[3]では情報拡散（ビデオ再生回数）には口コミ，ビデオの共有などの内因性（endogenous）の影響を受けた場合と，季節的なイベントや予期せぬ出来事による外因性（exogenous）の影響を受けた場合で分類されると述べている．更に他人が行動に影響を与える能力（subcritical - critical）で 4 つのパターン（endogenous-subcritical，endogenous-critical，exogenous-subcritical，exogenous-critical）に細分化される（図 2.2）と主張されている．ブログデータを用いた文献[4]では日々大量に発生する多様な情報のダイナミクスを分類化するために従来の K-means クラスタリングよりも効果的な K-Spectral Centroid クラスタリング手法を提案し，情報拡散過程を 6 つのパターンに分類化している（図 2.3）

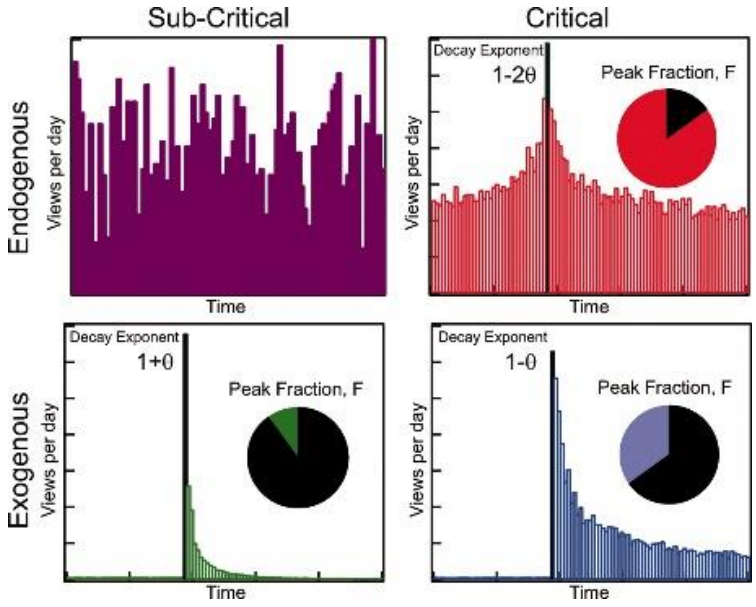

図 2.2：情報拡散過程の 4 パターン

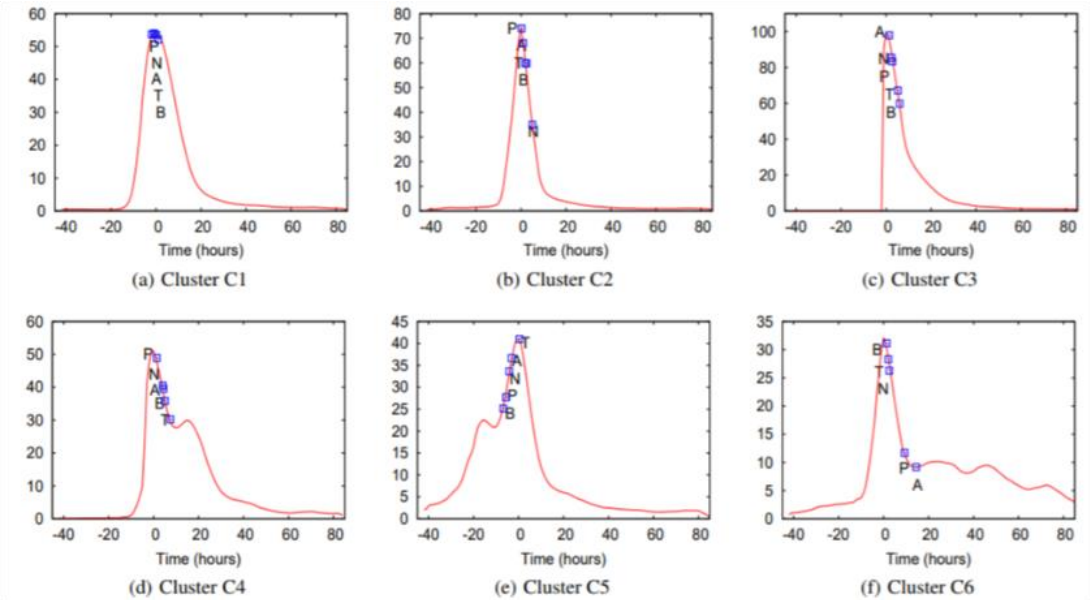

図 2.3：情報拡散過程の 6 パターン

2 つ目は情報拡散の拡散力を決定する重要人物の推定に関するものである．企業のマーケティング，災害時の情報共有など広範囲に隅々まで情報を拡散させたい場合にキーマンとなるユーザー（インフルエンサー）を特定することで効率的に拡散を行うことが出来る．

Cheong ら[5]の研究では 2010 年から 2011 年に発生したオーストラリアの洪水に関する Twitter 利用について実データを基に分析している．収集されたツイート数が 7000 と少数ではあるが，ツイートに対する返信がどのユーザーからされているかなどの関係をネットワークとして生成しネットワーク分析における様々な中心性の指標を導いている．分析の結果 Cheong らは地方自治体や国会議員，公共メディアが重要アカウントとして位置づけられていることを確認している．

石原らの研究[6]では Twitter ネットワーク全体を 1 つのネットワークとし，情報拡散と情報仲介の性質をネットワーク分析における次数中心性と媒介中心性の指標で表現することで重要人物を推定している．しかし Twitter ネットワークは趣味などによる情報嗜好の近い人間同士が繋がり合うものであると Amac らは述べている．[7]従って，推定された重要人物が一部のクラスタに偏っていた場合，情報拡散に偏りが生じる可能性がある．つまりネットワーク全体で重要人物の推定をすると，すべての人々に効率的に情報拡散出来る人物が必ずしも選択されるとは限らない問題がある．そこで，ページランクと大規模クラスタリングに基づく手法による重要人物の推定を行う研究がある．[8]

他にもツイートの内容に関する研究として風間ら[9]は東日本大震災でのツイート内容の分析を行っている．ツイートに出現する単語頻度の時間変化の類似性を Earth Mover's Distance(EMD)を用いて判定することで単語間の関連性の分析手法を提案している．この手法をスライディングウィンドウ方式により時系列分析することにより，地震，原発などに関連する単語の時系列変化を分析している．[10]

3 つ目はどのような情報が拡散されやすいかという情報の内容に関する研究である．Boyd らによるとリツイートされるツイートは時事的な速報ニュースであると述べられている．また，リツイートされるツイートは不安や怒りといったネガティブな内容のもの，個人的なものよりも公的，一般的な内容のものが多いということが示されている．
[11][12] [13]

第 3 章　ツイート数の時間変化

ここではキーワード検索により取得したツイートデータを用いて Twitter がリツイート率により 2 パターンの使われ方に分類できることを示す．さらに，1 分間当たりのツイート数，リツイート数，総ツイート数を実例とともに図で示す．なお審査委員からのコメントに基づき，データを追加して分析した結果についても図示する．

3.1　Twitter の持つ二面性

前章でも触れたが，SNS は個人のマイクロブログとしての使われ方に加えて，突出した出来事が発生した際の情報共有の場にもなるという二面性を持つようになってきている．ここでは得られた実際のツイートデータを基にその二面性の確認をする．

各キーワードで取得した際のリツイート数，総ツイート数，総ツイート数に占めるリツイート数の割合を表 3.1 に示す．日常的なキーワードで収集した際にはリツイート率が低く，災害や季節イベントなど突出した出来事に関するキーワードで収集した場合にはリツイート率が高いということが確認できる．ここから，災害などインパクトの大きい情報に関するツイートでは個人がツイートするよりもリツイートされやすい状態にあることが考えられる．

ある情報を受け取った人物が異なる人物にその情報を伝えることを情報共有と考えた時，Twitter のリツイートは情報共有そのものであり，得られたツイートデータからも二面性が確認できた．

表 3.1：キーワード毎のツイート数

キーワード	リツイート数	総ツイート数	リツイート割合	TYPE
嬉しい	83172	251998	33%	情報発信
悲しい	24416	110897	22%	情報発信
なう	134863	771402	17%	情報発信
インフル	302168	1045748	29%	情報発信
台風 24 号	1207848	1388199	87%	情報共有
シリア	81785	101318	81%	情報共有
中秋の名月	188141	244056	77%	情報共有
インドネシア	50308	75087	67%	情報共有
ゴーン	324039	418922	77%	情報共有
拡散希望	395688	538722	73%	情報共有
センター試験	585255	891442	66%	情報共有

3.2　情報共有された例

3.2.1　災害・事件

ここでは災害に関するキーワードでツイートを取得した.
1 例目は 2018 年 9 月 21 日に発生し 10 月 1 日に消滅した台風 24 号に関するツイートデータで，検索キーワードは「台風 24 号」，取得できたツイート数は約 139 万件である．取得したツイートの時間変化を図示する（図 3.1）

台風 24 号は全国 55 地点で最大瞬間風速が観測史上最大を更新する，JR 東日本が初の計画運休を実施するなど全国的に大きな被害を与える災害であったことから多くの人の関心を引いた事例である．台風発生の 1 日後からデータを取得した．発生直後ではツイート数は少なく 26 日頃から若干ツイート数に起伏が見られ始める．台風の接近とともにツイート数は増加していき，ツイート数のピークが一番初めに表れた日時は 29 日であった．この日は沖縄本島に上陸した日でもある．

最初のピーク終了は 30 日の午前 0 時頃であり，次のピーク時刻は 30 日の午前 7 時頃である．このように深夜時間はツイート数が大幅に減少する現象がみられた．これは深夜に Twitter のアクティブユーザー数が減少したことが原因として考えられる．台風 24 号に関して，全ツイート数（図中青線）のうちどの時間帯もリツイート数（図中緑線）が大部分を占める形となっている．なお，ツイート取得時，一番リツイートされたツイートは宇宙に関するツイートをつぶやくアカウント（フォロー数 12 フォロワー数 78373）で宇宙ステーションからみた台風 24 号の写真付きツイートであった.

一番リツイートされたツイートの 1 分間あたりのリツイート数を見てみる（図 3.2）
全体のツイート数と同じように夜間はリツイート数が少なく，朝から昼にかけてリツイート数が伸びていき，ピークを迎えた後減衰するという挙動をとる．図中の赤点はフォロワー数 100000 人以上のユーザーがリツイートした時間をプロットしたものである．このツイートを例にインフルエンサー（ここではフォロワー数が多いユーザー）と呼ばれるネットワーク上のハブにあたるユーザーの影響力を確認する．拡散されるツイートの特徴として，早期にインフルエンサーにリツイートされることが必要条件であるという研究がある [14]

しかしこの台風 24 号の例では，インフルエンサーがリツイートしている時にはすでにピークに到達しており，インフルエンサーが拡散現象を引き起こす割響を与えているとは考えにくい.

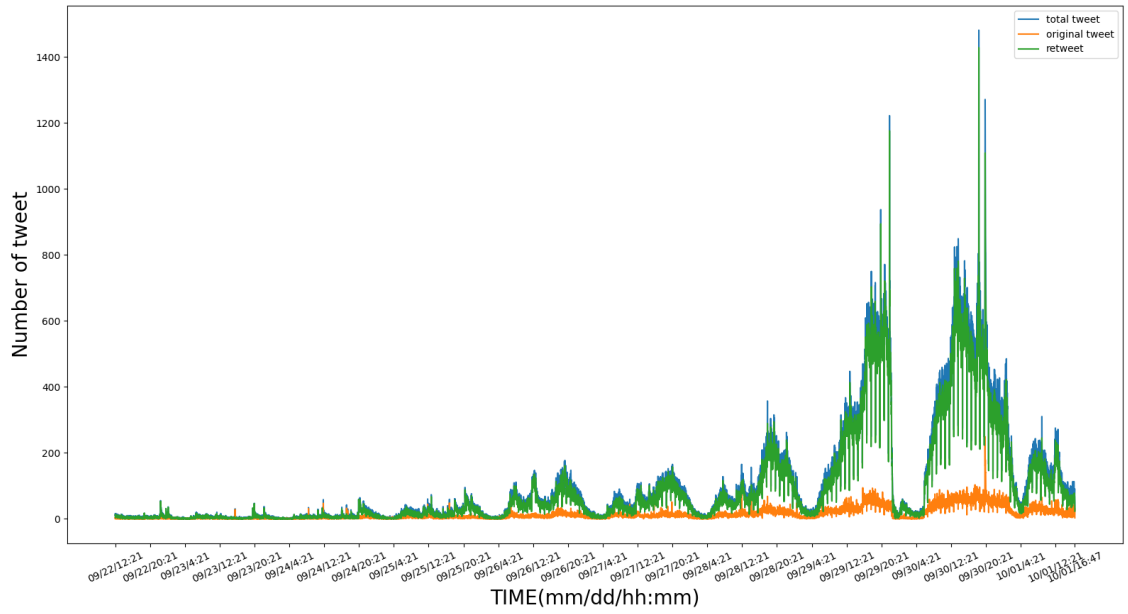

図 3.1 : 1 分間当たりのツイート数時間変化 (台風 24 号)

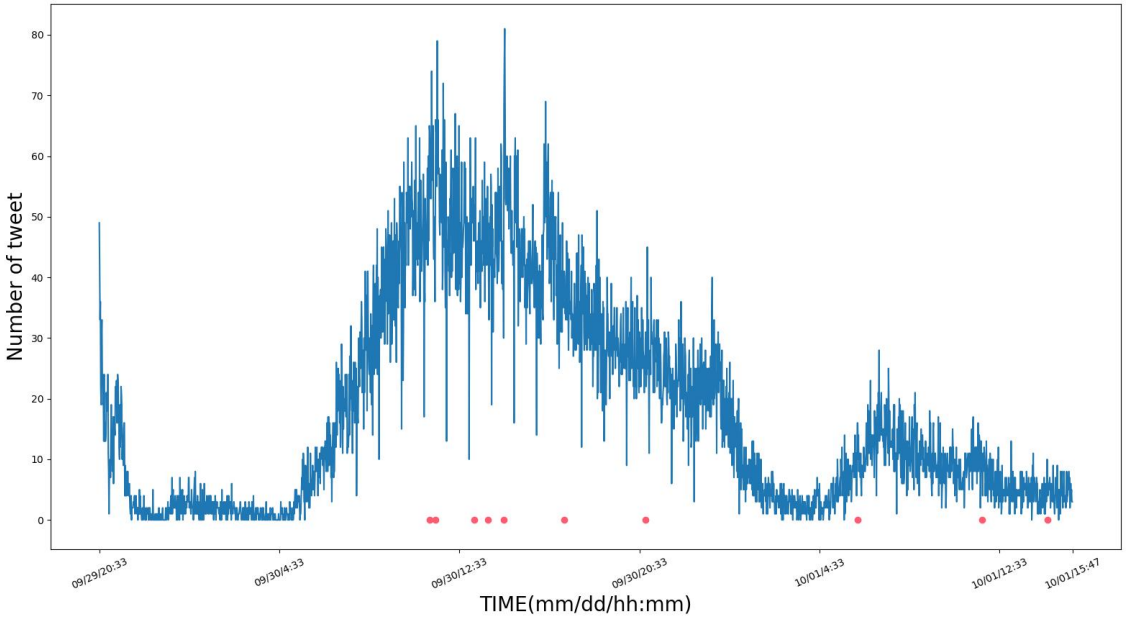

図 3.2 : 1 分間当たりのリツイート数時間変化 (台風 24 号)

2 例目は日本時間 2018 年 12 月 22 日 23 時 30 分にインドネシアで火山噴火による津波が発生し，死者 400 人超の災害が発生した．検索キーワードは「インドネシア」で取得件数は約 7 万 5000 件でリツイート率は 67%であった．1 分間あたりのツイート数を図 3.3 に示す.

特徴として災害発生前のインドネシアに関するツイートは少なかった.

また，災害発生直後の日本時間 22 日深夜から 23 日未明でも少ない．しかし 23 日朝方にかけて爆発的にツイート数が伸びており，多くの日本国民は災害翌日の朝に災害を認識し共有し始めたことが図示することで明らかになった.

3 例目は 2018 年 11 月 19 日に日産自動車会長のカルロスゴーン氏が金融商品取引法の疑いで逮捕された事件である．検索キーワードは「ゴーン」で取得ツイート数は 42 万件であった．ツイート数の 1 分間ごとの推移を図 3.4 に示す.

3 日間に渡る日数のツイートデータを取得したが，各日の正午にツイート数，リツイート数ともに増加していることが特徴として示された.

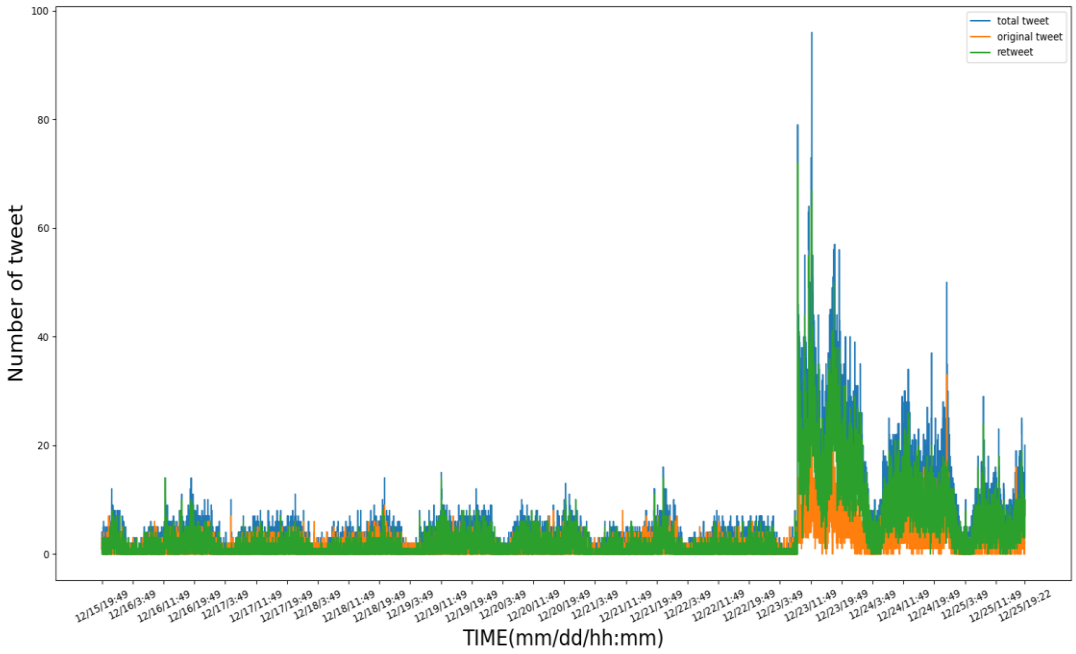

図 3.3：1 分間当たりのツイート数時間変化（インドネシア）

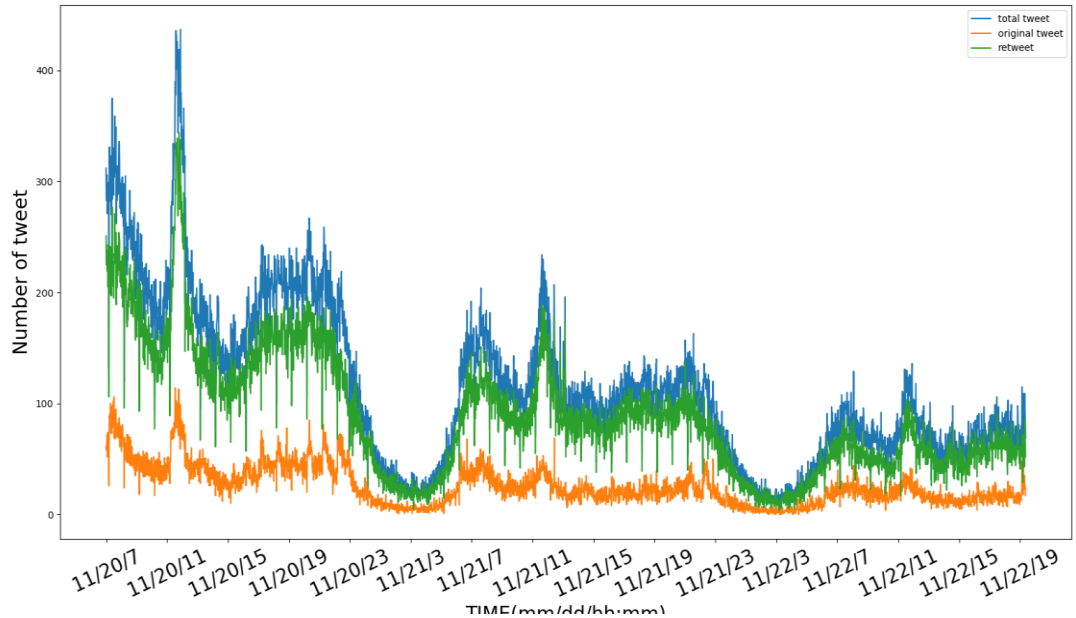

図 3.4：1 分間当たりのツイート数時間変化（ゴーン）

3.2.2　イベント

ここでは時期にまつわるイベントに関するキーワードでツイートを収集した．

1 例目は 2018 年 12 月 31 日に放映された紅白歌合戦に関するツイートである．キーワードは「紅白歌合戦」とした．1 分間あたりのツイート数推移を見てみると，31 日午前 11 時 10 分に急激にリツイート数が伸びていることがわかる（図 3.5）．このリツイート数の大部分は同時間帯に紅白歌合戦に出場していた米津玄師という歌手の公式アカウントからのツイートに対するもので，ツイート内容は紅白への感謝と来年の挨拶のみの簡潔なものであった．

同歌手の時間帯ではリツイート数は急激なピークを取るが，ツイート数にはピークは見られない．しかし紅白の放送終了時刻帯ではツイート数にも伸びが見られた．

Twitter 上で盛り上がりを見せているが，Twitter ネットワーク上のみに生じた特殊な盛り上がりの可能性がある．そこで，インテージ社が収集しているインターネットに接続された 80 万台超のテレビからの視聴状況を表したグラフ（図 3.6）と比較する．ここでは視聴率ではなく毎分のチャンネルの流入と流出の差で示される毎分接触率で紅白歌合戦の盛り上がりを確認する．ツイート数が急激に増加した要因である米津玄師が出演した時間帯の接触率を見ると，出演した時点をピークにツイート数と同様の急激な増減が見られる．このことから Twitter のみの盛り上がりではなく，世間的な盛り上がりを反映した結果であるといえる．

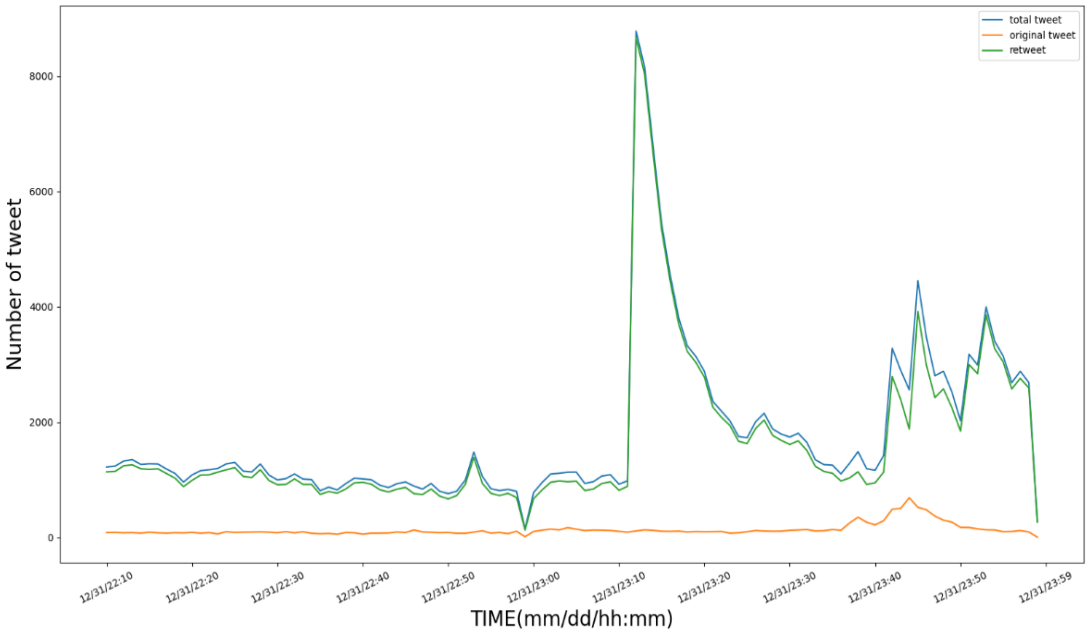

図 3.5：1 分間当たりのツイート数時間変化（紅白歌合戦）

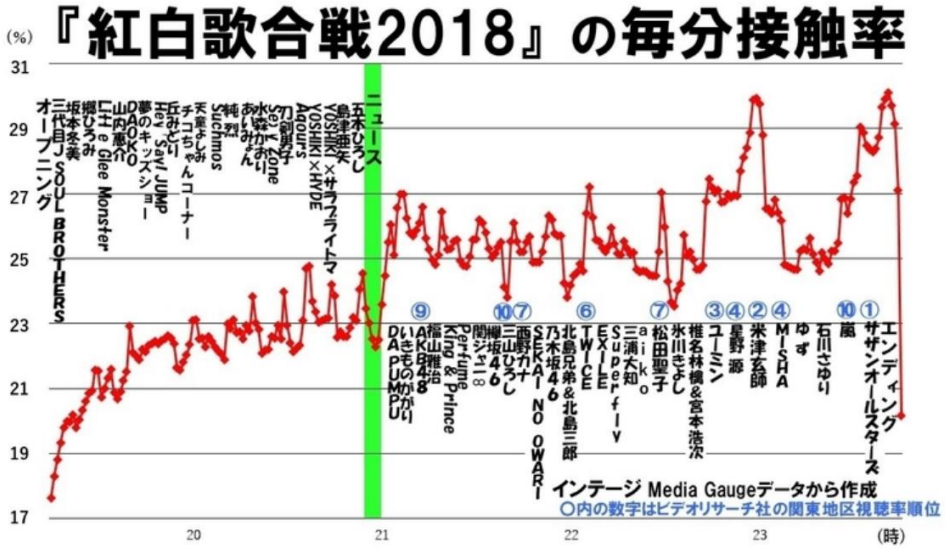

図 3.6：紅白歌合戦 2018 の毎分接触率

3.2.3　日常・個人

Twitter で情報共有を意図的に行う手段として「拡散希望」という言葉をツイートに含ませる方法がある．通常では情報発信に分類されるような個人的関心に基づくツイートが情報共有型になる例外をここでは取り上げる．

検索キーワードは「拡散希望」で取得ツイート数は 54 万件であった．

拡散希望で取得されたツイートの主な内容は個人的な情報の告知であり，社会に共通した関心事ではなかった．しかし図 3.7 からわかるようにリツイート数がツイート数よりも多い現象が発生している．

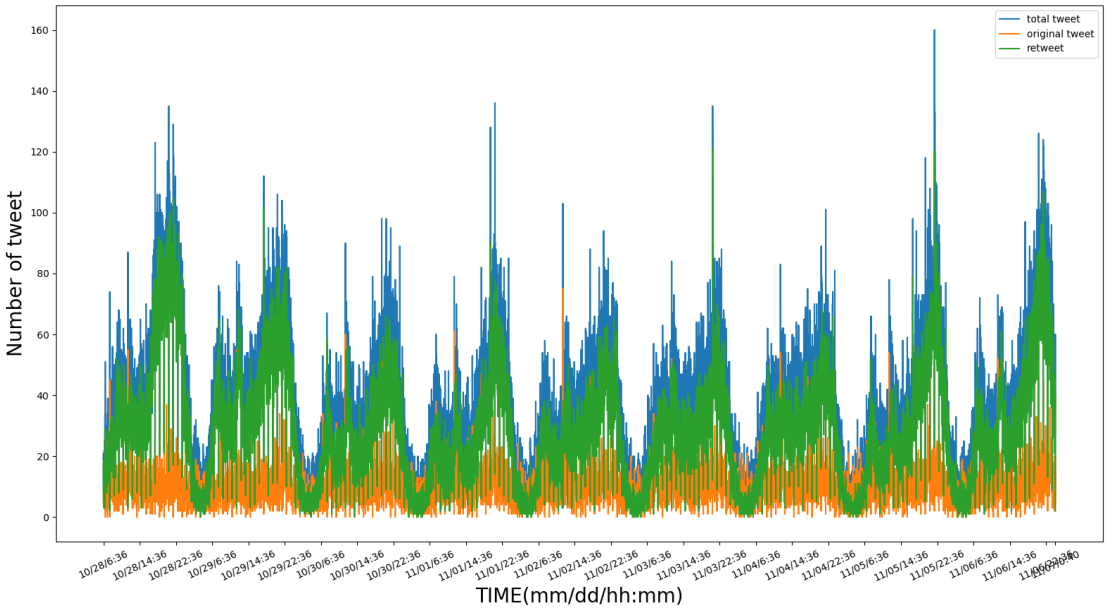

図 3.7：1 分間当たりのツイート数時間変化（拡散希望）

3.3 情報発信で使われた例

3.3.1 日常・個人

ここでは個人的，日常的な出来事についてツイートされたであろうものを収集した．
今していることを表す Twitter 独自に使用される単語でなうというものがある．
このキーワード「なう」で取得された 77 万件のツイート数の 1 分間あたりの時間変化を図示する（図 3.8）．

イベント，災害に関するキーワードで収集したツイートと違い規則的にツイート数が増減しており日付が変わってもツイート数，リツイート数ともに大きな変化は見られない．
また，イベントに関するキーワードで収集した場合，リツイート数が大半を占めるという結果を得られたが，この場合ではオリジナルツイート数が全ツイート数の大半を占めリツイート数は少数である．1 分間あたりのツイート数変化でみると確かにどの時間帯でもオリジナルツイート数（図中オレンジ線）がリツイート数（図中緑線）を上回っている．
「嬉しい」「した」など個人的な関心に基づくものは多くの人にとって関心のない情報であり，リツイートされにくい．

次にインフルエンザについてキーワード取得をした．検索キーワードは「インフル」で，約 105 万件収集されたツイートの時間変化について図示する（図 3.9）．
ここでも昼夜変動によるツイート数の増減が確認できる．リツイート数はどの時間帯でもツイート数を下回った．インフルエンザに関する話題から Twitter 上で多くの人々の関心を捉えたツイートは生まれなかったことが明らかになった．

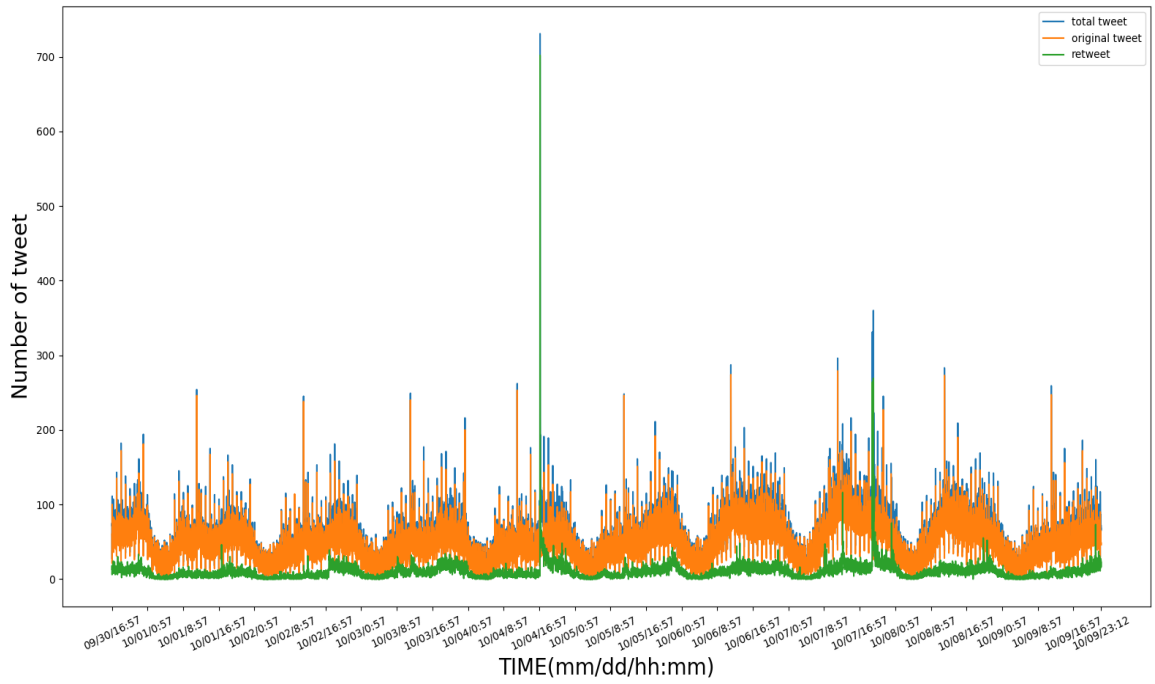

図 3.8：1 分間当たりのツイート数時間変化（なう）

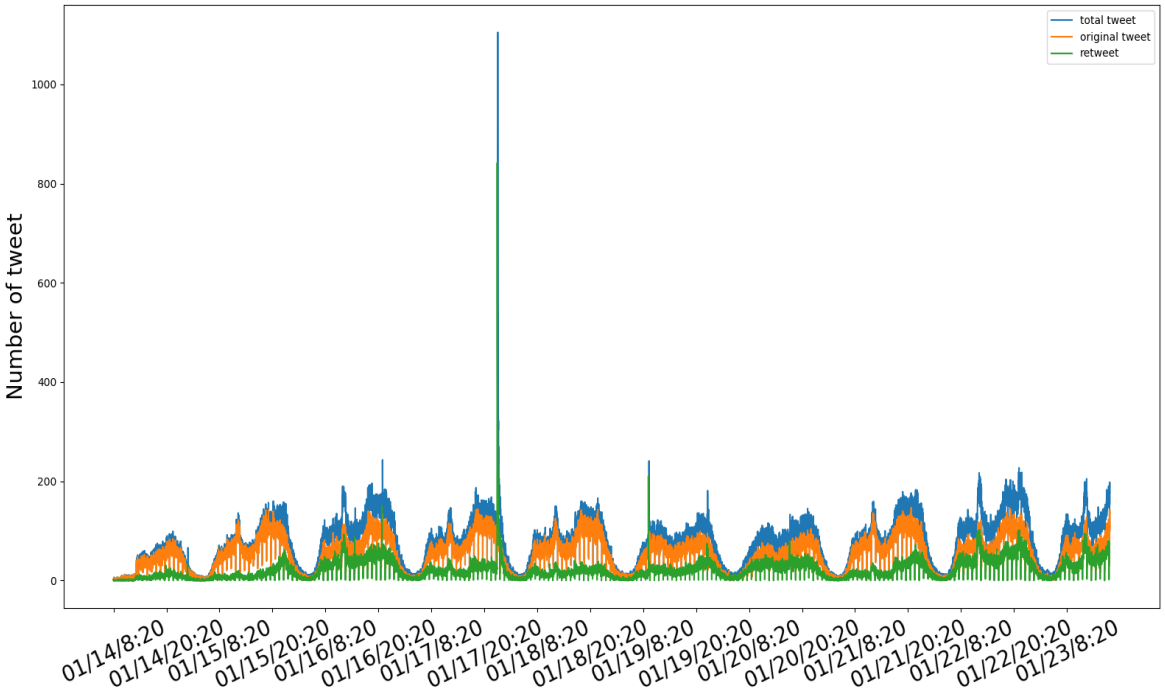

図 3.9：1 分間当たりのツイート数時間変化（インフル）

3.4　　情報発信・共有混合型

次に 2019 年 1 月 19 日 20 日に開催されたセンター試験についてツイートを収集した．キーワードを「センター試験」とし収集できたツイート数は 89 万件である．
図 3.10 より他キーワードで収集したツイートと同様に昼夜変動によるツイート数の増減が見られる．センター試験前日の 18 日昼から総ツイート数が急激に増加しており，ユーザーの関心の高まりがわかる．

このキーワードで収集したデータの特徴として，19 日以前と以後でリツイート数とツイート数の逆転が例外的に発生しており，前述した Twitter の 2 面性が同一キーワード内で見られる点である．
リツイート数の減少の方がツイート数の減少よりも緩やかに進行していることが原因と図から読み取れる．また，19 日以前ではセンター試験は実施されておらず．個々のユーザーの関心からなる事象であったものが，初日を終えた 19 日以降では共通した関心からなる事象へ変化したことにより，ツイートよりリツイートをするユーザーが増加したためである．

取得されたツイートデータの中で一番リツイートされたツイートの 1 分間あたりのツイート数時間変化を図示する．（図 3.11）このオリジナルツイートがツイートされた時刻は 19 日の 20 時 57 分でありツイート直後から爆発的にリツイートされている．ツイート内容はセンター試験のリスニング問題に登場したユニークな野菜キャラクターを 3D 作成した画像を添付したものであった．19 日はセンター試験にユニークなキャラクターが登場しネット上の話題となっていた．ユーザーの関心が高い状態が直後に膨大なリツイートを生んだと推測される．図中に示した赤丸はフォロワー数 10 万以上のインフルエンサーがリツイートした時刻であるが，台風 24 号の例と同様にインフルエンサーがリツイートのブームを作る現象はこの例でも確認できなかった．

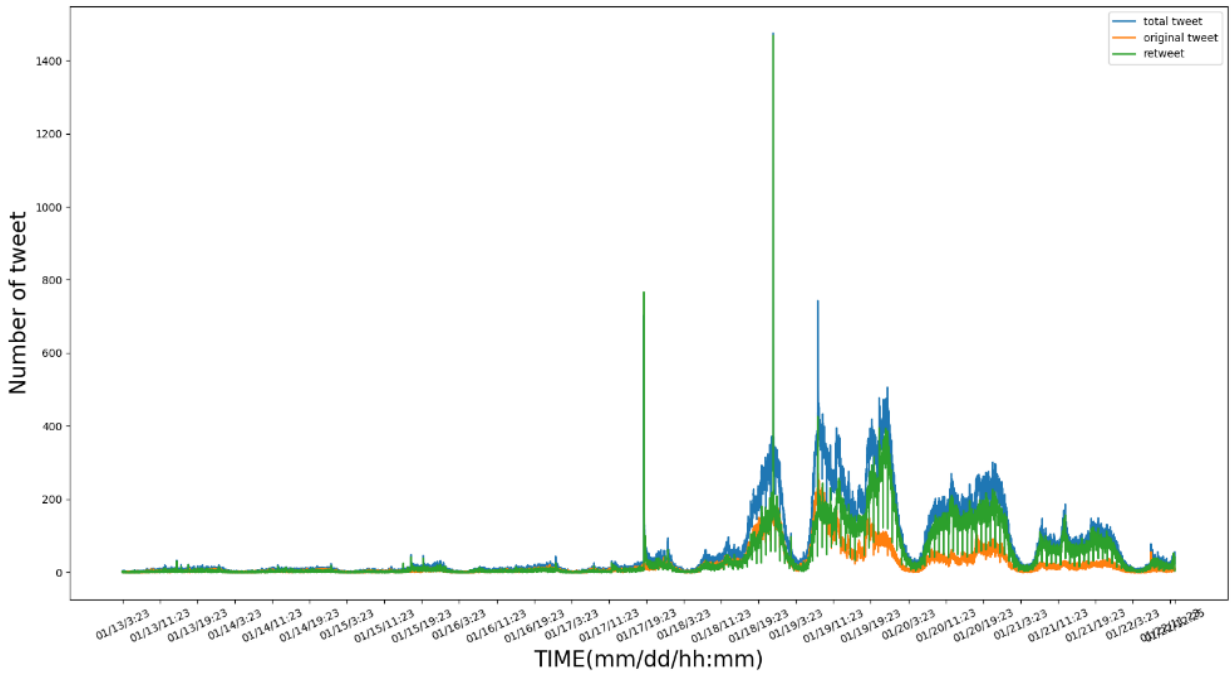

図 3.10 : 1 分間当たりのツイート数時間変化（センター試験）

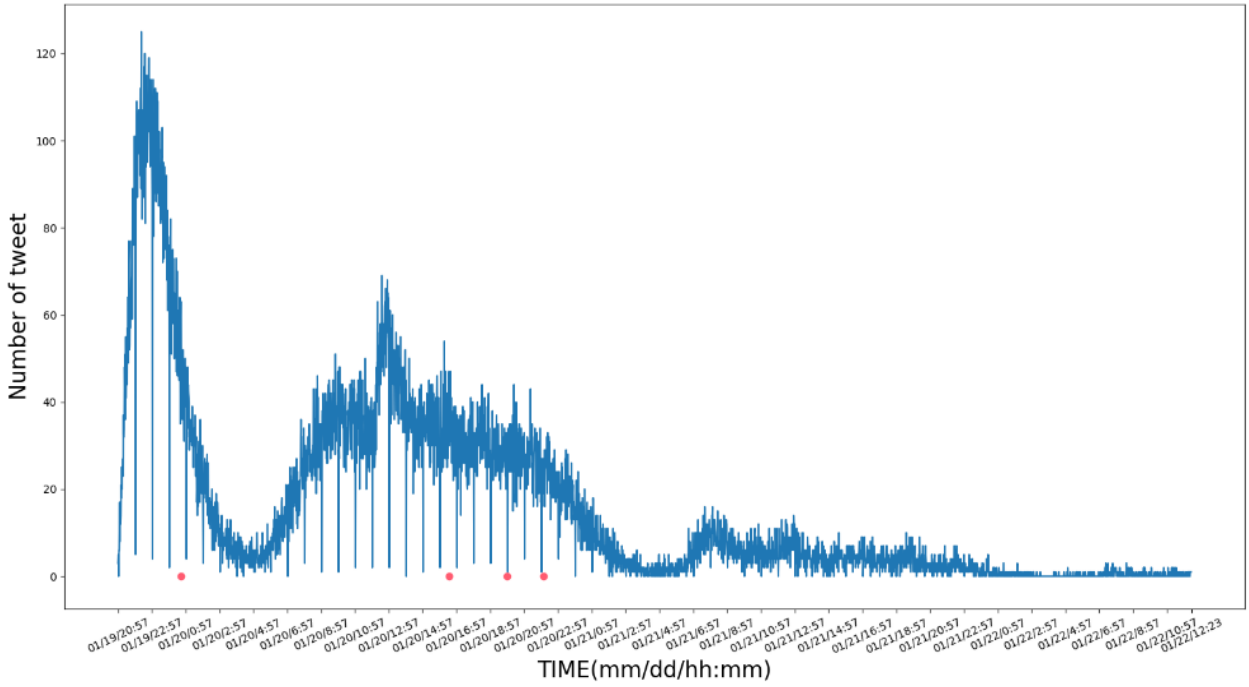

図 3.11 : 1 分間当たりのリツイート数時間変化（センター試験）

第 4 章　リツイート数の分布形状

本章では得られたツイートデータに共通してリツイート数の補分布を取ると裾の長い分布になることを明らかにする．また，各キーワードのリツイート数の両対数グラフを紹介する．

4.1　リツイート数の補分布

前章ではリツイート数の時間変化について示した．キーワード検索により取得したツイートデータ内にあるオリジナルツイートのリツイート数を確認しその分布を分析する．

各キーワードにおけるリツイート数の補分布を両対数グラフとして図 4.1～図 4.6 に示す．図の横軸をリツイート数(X)，縦軸はリツイート数がXを超えるツイートの割合（リツイート数の補分布）を示す．3.1 節で，Twitter は 2 面性を持つと述べたが，ここではいかなるキーワードで取得しても全てリツイート数の補分布は両対数グラフで直線状にプロットされ，べき分布を連想させる形状をとる．実際それぞれのツイートの最大リツイート数は平均リツイート数の数千倍を優に超える値をとっている．（表 4.1）また多くのキーワード検索で取得したツイートはリツイート数が 1 以下のツイートが総ツイート数の 80%以上を占めるという結果も得た．

両対数グラフでプロットした際の直線の傾きは 1 前後をとるものが多い．これはべき分布で近似した際にべき指数が 2 前後をとることを意味する．一般的にべき則で表される分布のべき指数γは$2 < \gamma < 3$で表すがそれとは対照的な特徴が見られる．

表 4.1　各キーワードの最大・平均リツイート数と傾き

キーワード	リツイート数平均値	リツイート数最大値	傾き
台風 24 号	9.078680	74135	-0.93292766
ハロウィーン	2.221010	54290	-0.92341024
なう	0.247939	8827	-1.08492059
拡散希望	2.993694	43893	-1.20599807
センター試験	3.0341	101984	-0.84532481
紅白歌合戦	15.886953	140749	-0.61992598
サンタ	1.492782	106365	-1.00632964
インフル	0.616848	52048	-0.75192932

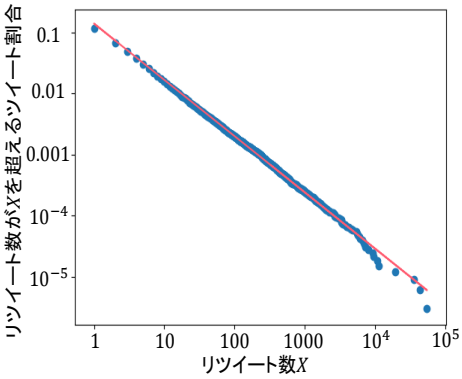

図 4.1：リツイート数の補分布
（ハロウィーン）

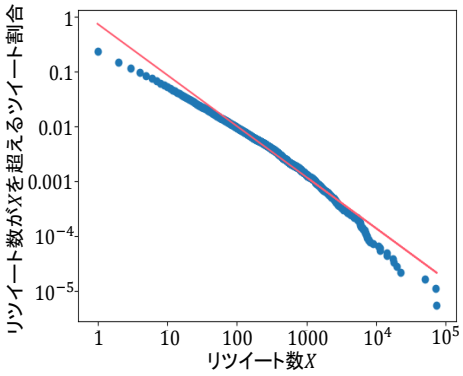

図 4.2：リツイート数の補分布
（台風 24 号）

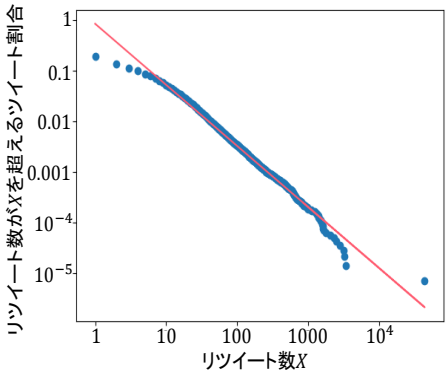

図 4.3：リツイート数の補分布
（なう）

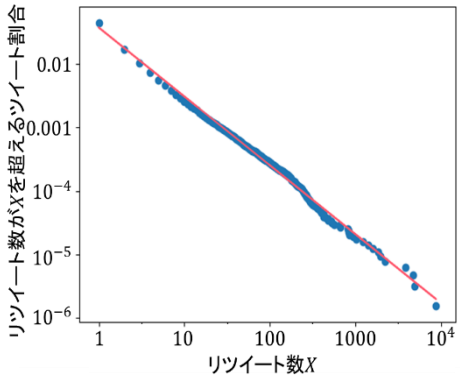

図 4.4：リツイート数の補分布
（拡散希望）

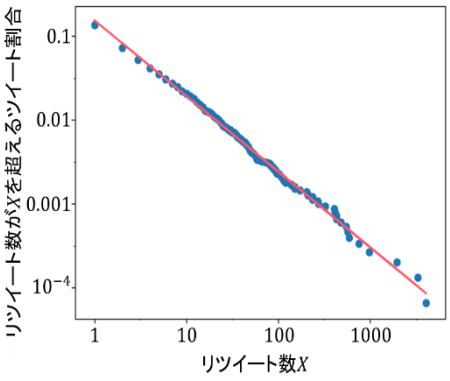

図 4.5：リツイート数の補分布
（日産）

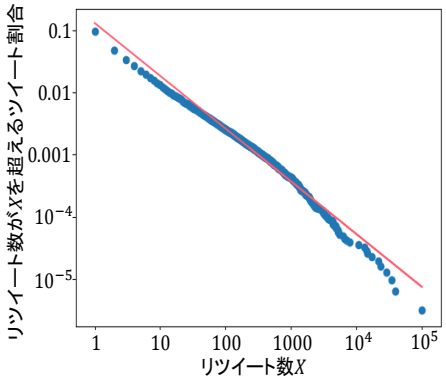

図 4.6：リツイート数の補分布
（センター試験）

4.2　　リツイート数と相関

前項でいかなるキーワードでもリツイート数にべき分布のような大きな偏りがあることを確認した.

ここでは，どのようなツイートがリツイートされているのか手がかりを探すために，リツイート数と相関関係にあるものを明らかにする．各キーワードのリツイート数とその相関を表 4.2 にまとめた.

リツイート数とオリジナルツイート主のフォロワー数については相関係数がどのキーワードで取得したものでも 0.2 以下であった．一般に相関係数が 0.2 を下回る場合，相関がないといえる．つまり，オリジナルツイートの投稿者とフォロー関係にあるユーザー数の大小は最終的にはリツイート数の大小に影響しないことが考えられる．次にリツイート数とツイート数の相関をみる．ここでいうツイート数とはオリジナルツイートを投稿したユーザーがアカウントを作成してからツイートした総数のことである．ここでも相関係数は 0.1 以下しか取らず，相関がないといえる.

最後にいいね数との相関を見る．いいねとはツイートに対するお気に入りを示すものである．いいね数とリツイート数は相関係数 0.9 ととても高い相関を示すことがわかった.

表 4.2　各キーワードのリツイート数とその相関

キーワード	対フォロワー数	対ツイート数	対いいね数
台風 24 号	0.08428848	-0.0014895	0.8830708
サンタ	0.08617871	-0.0002109	0.7257251
なう	0.19977945	0.0009469	0.9121638
ハロウィーン	0.10251513	0.0487327	0.7679788
高輪	0.06128336	0.0107373	0.9623689
中秋の名月	0.17487737	0.0007237	0.9671644
MondayMotivation	0.30087276	0.0278534	0.9102735
拡散希望	0.05048502	-0.0071356	0.9800733
センター試験	0.05311551	-0.0007007	0.96461948

第 5 章　　情報拡散モデル

本章ではリツイート数の時間推移を再現する簡単なシミュレーションモデルと数理モデルを提案する．

リツイート数の多いツイートでは，単位時間当たりのリツイート数推移は急峻なピークを迎えた後，昼夜変動を繰り返しながら減衰していくパターンをとっていた．ＳＮＳ上の情報拡散は様々なモデルが検討されている[15,16]が，ここではノード相関を考慮した情報拡散モデルを用いてその定型パターンが再現されるか検証する．

5.1　　モデル説明

先ほど説明した SIR モデルのとる 3 状態を SNS 上の情報拡散モデルでは以下の 4 状態をとることにする．

状態 0：ツイート（リツイート）を受け取っていない

状態 1：ツイート（リツイート）は受け取ったが，リツイートしない

状態 2：ツイート（リツイート）は受け取り，リツイート予定

状態 3：ツイート（リツイート）を受け取り，リツイート済み

また，Twitter を有向グラフで表現し，1 つ以上のノードを起点として有向グラフ内の各ノードへと拡散していく現象をモデル化する．ノードiは以下の状態を取る．

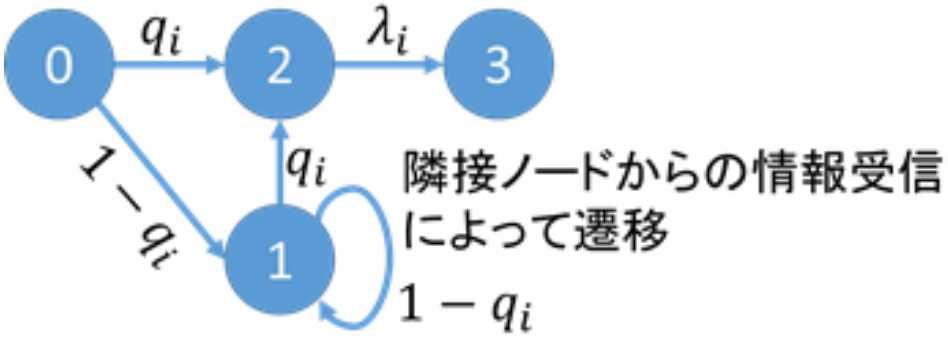

図 5.1:情報拡散モデルによる状態遷移図

初期状態ではツイート発信元のノードが状態 2 を取り，それ以外のノードは状態 0 をとる．ツイート発信元のノードが状態 2 から状態 3 へ遷移することで情報拡散が始まる．
リツイートを受け取っていない各ノードは隣接したノードからリツイートを受信することで状態 0 から確率qで状態 2 へ確率$1-q$で状態 1 へ遷移する．状態 2 に遷移したノードは，平均$1/\lambda$の指数分布に従う時間経過後に全ての隣接しているノードへ同時に転送（リツイート）し，状態 3 へと遷移する．状態 3 へ遷移後は状態 3 にとどまり，各ノードリツイートできる回数は 1 回のみとなっている．状態 0 から状態 1 へ遷移したノードはリツイートを受け取った別の隣接ノードから新たにリツイートされた場合，その時刻において確率qで状態 2 へ遷移する．全てのノードが状態 0，2，3 のいずれかになったとき拡散を終了する．なお，このモデルにおいて確率qはノードがツイート内容に興味を持つ確率として定義する．

時刻tにおいてノードiがとる状態を$Z_i(t)$で表すこととすると，ノード 1 からノードNまでで構成されるグラフの場合，時刻tでのネットワークの状態は各ノードの状態の組$(Z_1(t), Z_2(t), \dots, Z_N(t))$で記述される．ネットワークの状態遷移は連続時間マルコフ連鎖に従うので，シミュレーションにより情報拡散過程を追うことが可能である．

5.2　シミュレーション結果

情報拡散モデルを用いて条件を変化させたときのリツイート数の時間変化を図示する．各シミュレーション回数は 100 回行った．

5.2.1　発信ノードを変化させた場合

発信ノード次数によるリツイート数の時間変化をみる．発信ノード次数 1，10，100，1025（使用するネットワークの最大次数）を投稿主のフォロワー数としてリツイート数の時間変化を示す（図 5.2），確率$q = 1.0$，$\lambda = 1.0$とした．
発信ノード次数が 1 の時リツイート数のピークは他の場合と比べ遅く，小さいものになった．発信ノード次数が大きいとよりピークは早期に大きく出ることが判明した．前項でリツイート数とオリジナルツイートの投稿主のフォロワー数は相関が見られないことを述べたが，リツイート数の時間推移には影響があると考えられる．

5.2.2　平均時間λを変化させた場合

発信ノード次数 100，確率$q = 1.0$，λを 0.1，0.5，1.0，1.5 と変化させた場合について結果を示す．
λが最小の 0.1 の場合リツイート数は早期にピークを迎えピーク形状も急峻である．λが増加するに従い，ピークは緩やかに,遅れて訪れることを確認した．

5.2.3　確率qを変化させた場合

発信ノード次数 100，$\lambda = 1.0$，確率qを 0.1，0.5，1.0 と変化させた場合について結果を示す．

確率qを 0.1，0.5，1.0 と変化させた場合，総リツイート数は確率qの増加に比例して増えることを確認した．

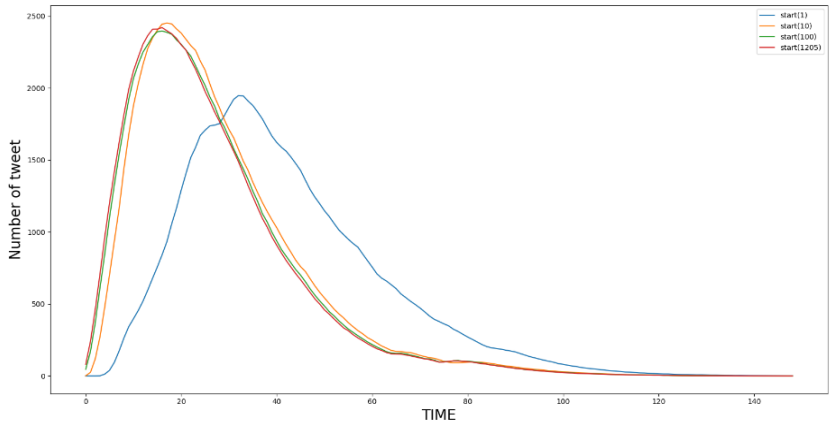

図 5.2 : 発信ノードを変化させた場合

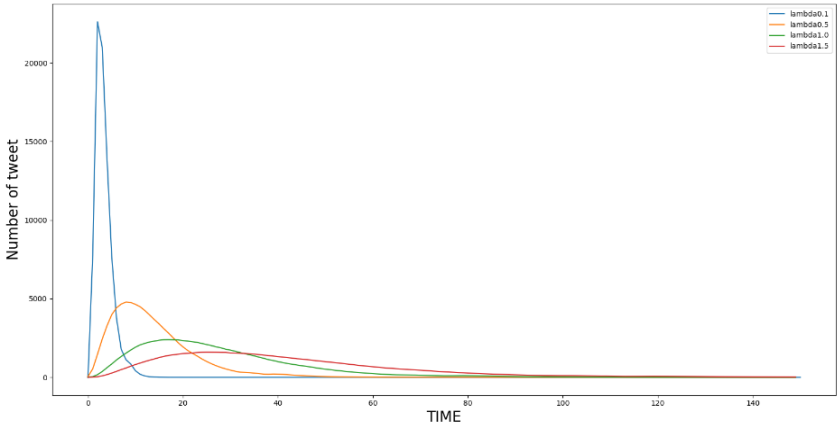

図 5.3 : 平均時間λを変化させた場合

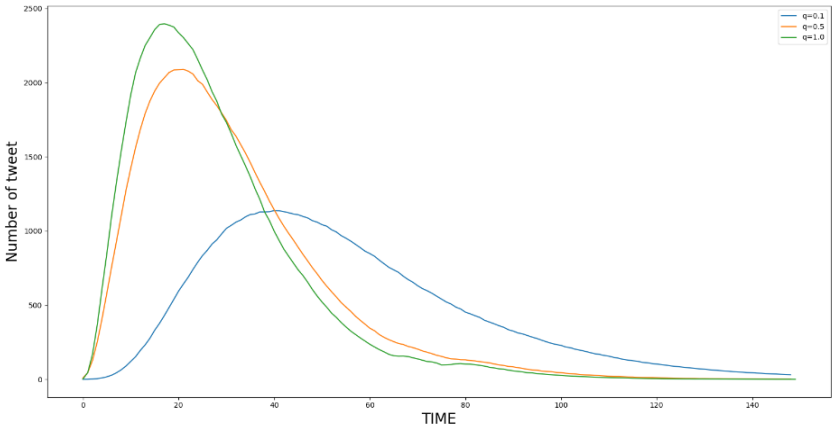

図 5.4 : 確率qを変化させた場合

5.3　　現実の事象の再現性

前項でパラメーターλ，qを変化させた場合のリツイート数推移の挙動を参考に実際に取得したツイートの再現をする．

前節のモデルはネットワーク構造，λ，qの 2 つのパラメーターを含む．現実の Twitter を構成するネットワークトポロジーを入手することは困難であるため，ネットワーク構造については仮定を置くものとして，2 変数λ，qに適当な値を取ることで現実のリツイート数の再現が可能になると考える．前節のモデルでは昼間はリツイート数が増え，深夜はリツイート数が落ち込むといった時間変動を考慮していないが，ノードがリツイートするまでに経過する時間λを時間に関する関数とすることで時間変動を取り入れる．各変数値を表 5.1 にまとめる．

ここで使用する Twitter トポロジーデータはノード数 81306，リンク数 1768149 である．[17]リツイートは各ノード 1 回しか出来ないので今回使用するトポロジーでは，総リツイート数が 81306 回を超える再現はできない．なお，81306 回を再現する際には全てのノードが確率$q = 1.0$で興味を持ち，必ずリツイートするという条件を取っている．

以上の前提条件を踏まえた上で再現した結果を示す．

図 5.5 はセンター試験で取得したツイートの中で一番リツイートされたツイートの再現である．青線がモデルによる再現，橙線が実データによるものである．

ツイート直後に最大のピークをとり夜間にツイート数は落ち込み，翌朝にまたリツイートされ始め 2 度目のピークを取るという現象を再現できた．しかし実データで示された 3 度目のわずかなピークは再現できなかった．

次に紅白歌合戦で最もリツイートされた物の時間変化を再現したものを図示する．（図 5.6）こちらは先ほどの例とは異なり 1 時間という短時間で大きな推移をしたものの再現である．ツイート直後に最大のピークを迎え，急速に減少していく様子を再現できている．

表 5.1　各キーワード再現時の 2 変数の値

	λ(0-10)	λ10-20	λ深夜	λその他	確率q
センター試験	0.13	1	1.3	3	0.3
紅白	0.06	0.09	0.4	0.4	0.8

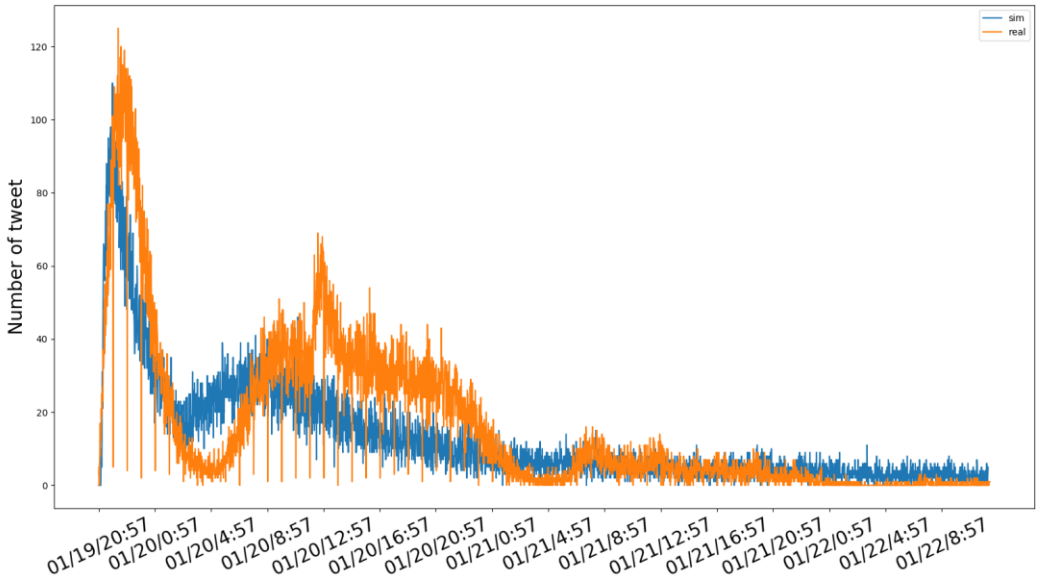

図 5.5：リツイート数の時間変化再現

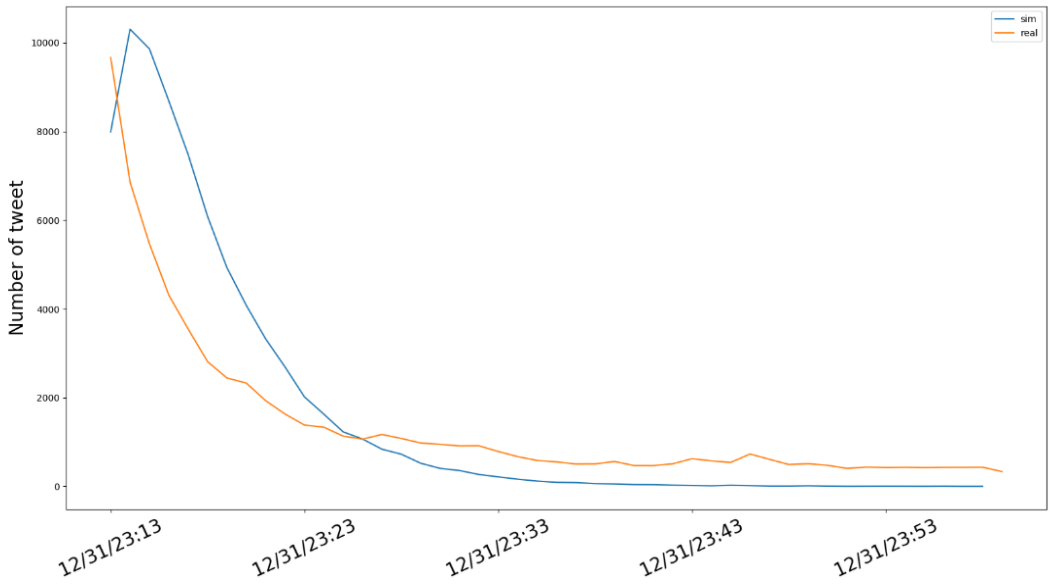

図 5.6：リツイート数の時間変化再現
（紅白歌合戦）

5.4　　数理モデル

ここでは，情報拡散過程について SIR モデルを参考に数理モデルを提案し，シミュレーションを行わずに表現する．また，数理モデルを導出する際に隣接した 2 つのノードの相関性についての扱いについて述べる．

5.4.1　　SIR モデル

数理モデルを構築する際に参考とした SIR モデルについて簡単に説明する．
SIR モデルとは感染症の流行過程を記述するシンプルなモデルで，健康状態(Susceptible:状態 S）のノードが感染状態（Infected:状態 I）の隣接ノードから確率的に感染し状態 I へ遷移する．病気が治癒すると免疫獲得状態（Recovered:状態 R）に遷移し再び感染することはなくなるといったモデルである（図 5.7）．

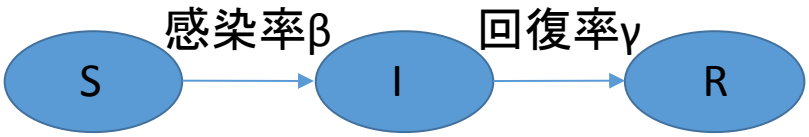

図 5.7：SIR モデルの状態遷移

5.4.2　　モデル説明

ここでは，ノードの取る状態を以下の 3 状態に定義する．
状態 0：ツイート（リツイート）を受け取っていない
状態 1：ツイート（リツイート）を受け取り，リツイート予定
状態 2：ツイート（リツイート）を受け取り，リツイート済み
ノード状態の遷移は状態 0 のノードはリツイートを受け取ると状態 1 へ遷移する．状態 1 へ遷移したノードは指数分布に従う時間経過後に全ての隣接ノードへ同時にリツイートを行い，状態 2 へ遷移する．なお，状態 2 への遷移後はリツイートを行わず状態 2 にとどまるものとする．

数理モデルでは，隣接行列と各ノードの状態 1 から状態 2 への遷移率をパラメーターとして有する．以下では，隣接行列を$A = \{a_{ij}\}$(a_{ij}はノードiからノードjへの有向リンクが存在した場合 1，しない場合は 0 をとる変数)，ノードiの状態 1 から状態 2 への遷移率をλ_iとする．

5.4.3　　解析手法

本節では解析的に分析することを試みる．ネットワークの状態遷移を記述するマルコフ連鎖が既約で正再帰的であれば，全てのノードが状態 2 へ遷移し定常状態となる．本モデルでは定常状態は自明であるので，初期状態から定常状態への遷移過程（過度特性）を解析する．しかし，ネットワーク状態数（3^N）はノードNの増加とともに爆発的に増大するので，過度特性を厳密に解析することは不可能である．

そこで，ノードiが時刻tで状態kにいる確率を$p_i^{(k)}$で表す．$p_i^{(k)}$の時間変化に関する微分方程式を数値的に解くことで過度特性をここでは評価する．このため，以下の確率変数を導入する．

$$X_i^{(k)}(t) = \begin{cases} 1, & Z_i(t) = k \\ 0. & otherwise \end{cases}$$

このとき，ノードiが時刻tで状態kにいる確率$p_i^{(k)}$は以下のように表される．

$$p_i^{(k)}(t) = E\left[X_i^{(k)}(t)\right]$$

ノードiが状態 1 へ遷移するにはノードjからノードiへの有向リンクがあり，ノードjが状態 1 かつノードiが状態 0 の時である．ノードiの状態 1 から状態 2 への遷移は遷移率λ_iのみで決まり，状態 2 で留まることから，

$$\frac{dp_i^{(1)}(t)}{dt} = -\lambda_i E\left[X_i^{(1)}\right] + \sum_j a_{ji}\,\lambda_j E\left[X_j^{(1)}(t)X_i^{(0)}(t)\right] = -\lambda_i p_i^{(1)}(t) + \sum_j a_{ji}\,\lambda_j E\left[X_j^{(1)}(t)X_i^{(0)}(t)\right]$$

$$\frac{dp_i^{(2)}(t)}{dt} = \lambda_i E\left[X_i^{(1)}\right] = \lambda_i p_i^{(1)}(t)$$

ここで，$X_i^0 + X_i^1 + X_i^2 = 1$であることから$X_i(\mathrm{t}) \stackrel{\text{def}}{=} \mathrm{X_i^{(1)}} + \mathrm{X_i^{(2)}}$，$p_i(t)$を時刻$t$でノード$i$がリツイートを受け取る確率とおくと，$p_i(t) \stackrel{\text{def}}{=} p_i^{(1)} + p_i^{(2)}$であることから，

$$\frac{dp_i(t)}{dt} = \sum_j a_{ji}\lambda_j E\left[X_j^{(1)}(t)(1 - X_i(t))\right]$$

を得る．

さらに，ノードjがリツイート済みならばノードiは必ずツイートを受け取っているので．ノードiの状態は 1 か 2 である．つまり$a_{ji} = 1$ならば，$X_j^{(2)}(t) = 1$のとき$\mathrm{X_i}(t) = 1$である．よって$a_{ji}X_j^{(2)}(t)\big(1 - X_i(t)\big) = 0$であり，

$$\begin{aligned} \frac{dp_{i(t)}}{dt} &= \sum_j a_{ji}\lambda_j E\left[X_j^{(1)}(t)(1 - X_i(t))\right] \\ &= \sum_j a_{ji}\lambda_j E\left[(X_j^{(1)}(t) + X_j^{(2)}(t))(1 - X_i(t))\right] \\ &= \sum_j a_{ji}\lambda_j E\big[X_j(t)(1 - X_i(t))\big] \\ &= \sum_j a_{ji}\lambda_j (p_j(t) - E\big[X_j(t)X_i(t)\big]) \end{aligned} \qquad (1)$$

を得る．

時刻tまでに行われたリツイート回数を$L(t)$とすると

$$L(t) = \sum_i X_i^{(2)}(t)$$

であるからその期待値は

$$E[L(t)] = \sum_i E\left[X_i^{(2)}(t)\right] = \sum_i p_i^{(2)}(t)$$

である．

ここでノードiがリツイートを受信した時刻をT_i，リツイートを受信してからリツイートするまでに経過した時間をτ_iとすると

$$\begin{aligned} p_i^{(2)}(t) &= P(T_i + \tau_i \le t) \\ &= \int_0^t P(T_i \le t - s)P(s \le \tau_i < s + ds) \\ &= \int_0^t \lambda_i P(T_i \le t - s)e^{-\lambda_i s} ds \\ &= \int_0^t \lambda_i p_i(t - s)e^{-\lambda_i s} ds = \int_0^t \lambda_i p_i(s)e^{-\lambda_i (t-s)} ds \end{aligned}$$

従って

$$E[L(t)] = \sum_i \int_0^t \lambda_i p_i(s)e^{-\lambda_i (t-s)} ds \qquad (2)$$

式（1）について数値的に解き$p_i(t)$を求めることで（2）からリツイート数の期待値の時間変化を計算することが可能である．

微分方程式（1）を解く際，（1）右辺に現れる$E\left[X_j(t)X_i(t)\right]$は既知でないので（1）から$p_i(t)$を数値的に計算するために，仮定を置いて$E\left[X_j(t)X_i(t)\right]$を$p_i = E[X_i(t)]$と$p_j = E[X_j(t)]$の関数で表現する必要がある．以下$E\left[X_j(t)X_i(t)\right]$を相関項と呼ぶこととする．

5.4.4　独立モデル

微分方程式（1）を解く際に必要となる相関項の扱いについて述べる．

相関項を各ノードの積で仮定し近似（$E\left[X_iX_j\right] \approx E[X_i]E[X_j]$）する手法を独立モデルとよぶ．[R.Pastor-Satorrar,C. Castellano,P.V.Mieghem,ansA.Vespignani,“Epidemic processes in complex networks,”Reviews of Modern Physics,vol.87,pp.926-979,2015]

独立モデルでは式（1）は以下のように表される．

$$\frac{dp_{i(t)}}{dt} = \sum_j a_{ji}\lambda_j(p_j(t) - E[X_j(t)X_i(t)])$$
$$= \lambda_j \sum_j a_{ij}p_j(t)(1 - p_i(t))$$

5.4.5 強相関モデル

ここではノードiとノードjが独立でない場合を考える.
$X_i(t)$と$X_j(t)$が非負の相関を持つことから,

$$\mathrm{Cov}[\mathrm{X_i}(t), X_j(t)] = E[X_i(t)X_j(t)] - E[X_i(t)]E[X_j(t)] \geq 0$$

従って

$$p_i(t)p_j(t) \leq E[X_i(t)X_j(t)]$$

また,

$$E[X_i(t)X_j(t)] = P(\{X_i(t) = 1\} \cap \{X_j(t) = 1\})$$
$$\leq P(\{X_i(t) = 1\}) = p_i(t)$$

および

$$E[X_i(t)X_j(t)] = P(\{X_i(t) = 1\} \cap \{X_j(t) = 1\})$$
$$\leq P(\{X_j(t) = 1\}) = p_j(t)$$

であることから

$$E[X_i(t)X_j(t)] \leq min\{p_i(t), p_j(t)\}$$

が成り立つ.
ここでは,

$$E[X_i(t)X_j(t)] = min\{p_i(t), p_j(t)\}$$

とする近似を強相関近似と呼び, この近似をとるモデルを強相関モデルと呼ぶ.
強相関モデルでは式（1）は以下のように表せられる.

$$\frac{dp_{i(t)}}{dt} = \sum_j a_{ji}\lambda_j(p_j(t) - E[X_j(t)X_i(t)])$$
$$= \sum_j a_{ji}\lambda_j(p_j(t) - min\{p_i(t), p_j(t)\})$$
$$= \sum_{j;p_j(t)>p_i(t)} a_{ji}\lambda_j(p_j(t) - p_i(t))$$

第 6 章　リツイート数のべき則性出現モデル

前述したように検索キーワードによらずリツイート数は共通してべき分布のような裾の長い分布に従うことがわかった．
本章では， Barabasi-Albert モデルの優先的選択ルールを参考にリツイートモデルを提案しべき分布に従う現象を再現する．

6.1　Barabasi-Albert モデル

Barabasi-Albert モデル[18]とは，スケールフリー性を実現するネットワークモデルの 1 つである．ノードは次々とネットワークに加わるというネットワークの成長を考える．成長の際新しく加わったノードは既存のノードと等確率でリンクを張らず，次数の高いノードとリンクを張りやすいとする．このような成長と優先的選択という仕組みを持ったネットワーク形成モデルである．

6.2　モデル説明

Barabasi-Albert モデルを参考にリツイート数のべき則性を再現するモデルを提案する．
ここでは簡単のために，ネットワークトポロジーは完全グラフとする．
時刻tまでに書き込まれたツイート総数を$N_0(t)$，時刻tまでにリツイートされた総数を$N_1(t)$，n番目に書き込まれたツイートの時刻tでのリツイート数を$r_n(t)$，n番目に書き込まれたツイート内容を点数化したものをa_nとする．
以下の要領でモデル化する．

(1) 時刻 0 以降，頻度λ_0でツイートが書き込まれる．
(2) 時刻 0 以降，頻度λ_1でその時刻までに書き込まれた全ツイートの中から 1 つツイートが選択されリツイートされる．
(3) (2)において，n番目に書き込まれたツイートは確率$\frac{(a_n+r_n(t))}{\sum_{i=1}^{N_0(t)}(a_n+r_i(t))}$で選択される．

a_nはn番目のツイートの価値（面白さ）を数値化したものに相当する．

簡単のため，以下$a_1 = a_2 = \cdots = a$とする．

リツイートされる機会は単位時間当たりλ_1回存在し,1 回あたりn番目に書き込まれたツイートがリツイートされる確率は$(a_n + r_n(t))/\sum_{i=1}^{N_0(t)}(a_n + r_i(t))$であることから，次が成り立つ.

$$\frac{dr_n(t)}{dt} = \lambda_1 \frac{a + r_n(t)}{\sum_{i=1}^{N_0(t)}(a + r_i(t))}$$

ここで，時刻tまでに書き込まれたツイート総数はおよそ$\lambda_0 t$に等しく，同時刻までのリツイート総数はおよそ$\lambda_1 t$に等しいことより

$$\sum_{i=1}^{N_0(t)} \big(a + r_i(t)\big) = aN_0(t) + N_1(t) \approx (a\lambda_0 + \lambda_1)t \qquad (1)$$

n番目に書き込まれたツイートの投稿時刻をt_nとする. $r_n(t_n) = 0$の初期条件で(1)を解くと，次を得る.

$$r_n(t) = a\left(\left(\frac{t}{t_n}\right)^{1/\gamma} - 1\right),\ \gamma \stackrel{\text{def}}{=} \frac{a\lambda_0 + \lambda_1}{\lambda_1}$$

上式より，n番目に書き込まれたツイートのリツイート数がxを超える，$a\left(\left(\frac{t}{t_n}\right)^{1/\gamma} - 1\right) > x$であることは,

$$t_n < t\left(\frac{a}{a + x}\right)^{\gamma}$$

であり，そのツイートが時刻$t\left(\frac{a}{a+x}\right)^{\gamma}$以前に書き込まれたことを意味する．ツイートは一定頻度で書き込まれているので時刻tまでに書き込まれたツイートのうち, 時刻$t\left(\frac{a}{a+x}\right)^{\gamma}$以前に書き込まれたツイートの割合は$\left(\frac{a}{a+x}\right)^{\gamma}$に等しい．従ってリツイート数$x$を超えるツイートの割合$P(r > x)$は

$$P(r > x) = \left(\frac{a}{a + x}\right)^{\gamma} \approx x^{-\gamma}$$

つまり，リツイート数の分布はべき則に従い，そのべき指数は$\gamma + 1$と等しい．このモデルでは$\gamma \geq 1$である.
例として，図にツイート総数n，リツイート頻度$\lambda_1 = 1$，ツイート頻度$\lambda_0 = 3$の場合を図示する.

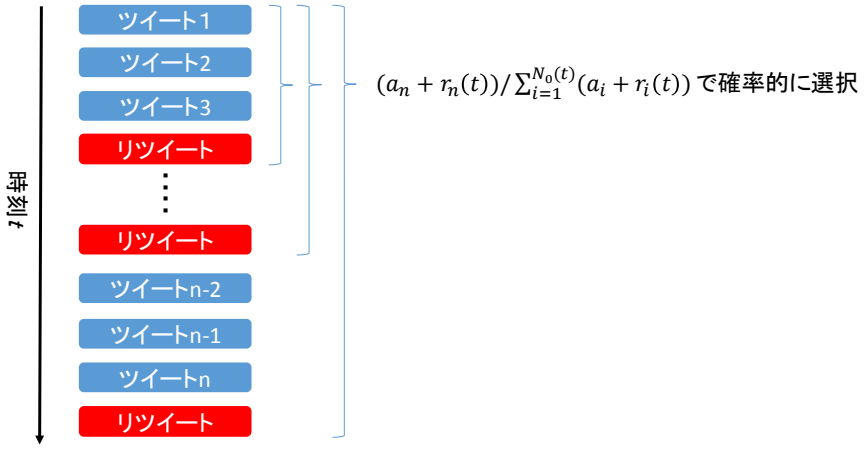

図 6.1 : リツイート数のべき出現モデル

なお，このモデルは$a \leq 0$とすると 1 番目のツイートしかリツイートされなくなるが，リツイートモデルの 3 番目の項目を以下に変えると$a \leq 0$の条件も満たし，γが 1 未満（べき指数 2 未満）の分布が再現できる．

3* 2 において，n番目に書き込まれたツイートは確率
a_nはn番目のツイートの価値（面白さ）を数値化したものに相当する．

4.1 節の結果からリツイート数の補分布はγ ≤ 1（べき指数 2 以下）である．
これは$a \leq 0$の時であり，$a \leq 0$はリツイート数が$|a|$を超えるツイートのみリツイートの対象とすることを意味する．
つまり，Twitter 上で情報拡散を行う際，人々はツイートの価値ではなくリツイート回数に基づいてリツイート行動を起こしていることを表す．

6.3 現実の事象の再現性

ここでは，提案モデルを用いてシミュレーションにより各ツイートのリツイート数を再現する．図 6.2 はツイート頻度$\lambda_0 = 1$，リツイート頻度 $\lambda_1 = 10$，a_nを 0 から 5 までのランダムな数値をおいて提案モデルによるシミュレーションを行った結果である．また，図中青点で示しているのは「サンタ」というキーワードで取得したツイートのリツイート数の補分布である．図 6.3 は$\lambda_0 = 1, \lambda_1 = 2, a_n$を 0 から 1 までのランダムな数値でおいた場合で青点は「Monday Motivation」というキーワードで取得したツイートのリツイート数の補分布を示す．図 6.4 は$\lambda_0 = 2, \lambda_1 = 12, a_n$を 0 から 0.6 までのランダムな数値でおいた場合であり，青点は「センター試験」で得られたリツイート数の補分布である．
図 6.5 は$\lambda_0 = 3, \lambda_1 = 1, a_n$を 0 から 1 までの値で，青点は「日産」で得られたツイートのリツイートの補分布である．
提案モデルによるリツイート数の補分布が現実の事象を再現できていることが確認された．

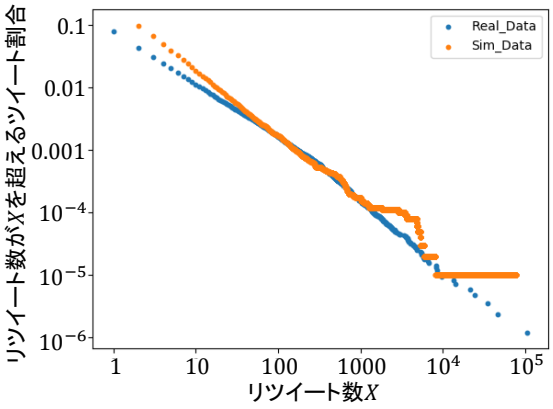

図 6.2：提案モデルによるリツイート数の補分布と実データ（サンタ）

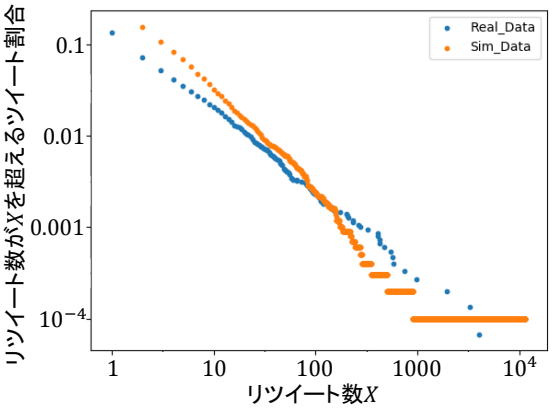

図 6.3：提案モデルによるリツイート数の補分布と実データ（MondayMotivation）

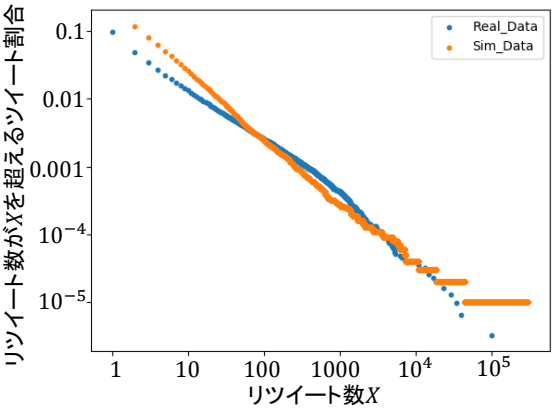

図 6.4：提案モデルによるリツイート数の補分布と実データ（センター試験）

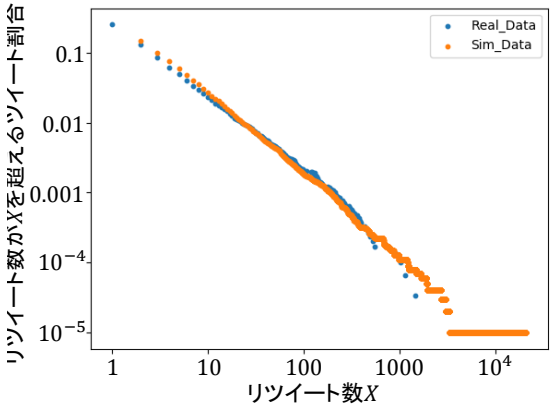

図 6.5：提案モデルによるリツイート数の補分布と実データ（日産）

6.4　　シミュレーションによる相関

前項では実データを用いてリツイート数といいね数やツイート数の相関をみた．本項では提案したモデルを元にリツイート数をシミュレーションし，リツイート数とツイート順の相関，リツイート数とツイートの点数（内容）との相関を見る．

6.4.1　　リツイート数とツイートの面白さの相関

リツイート数とツイートの内容の相関を各シミュレーション条件（リツイート頻度とツイート内容を点数化する際に点数がとりうる値の範囲）でそれぞれまとめたものが表 6.1 である．なおツイート頻度は 1 とし，() 内の数値はシミュレーション結果を両対数グラフで表した際の傾きである．

表 6.1 からリツイート頻度が増えるにつれ，リツイート数とツイートの面白さの相関は減少する傾向がある．また各ツイートの点数に幅を持たせた場合，同じリツイート頻度でもリツイート数とツイートの点数との相関は高くなることが判明した．ツイートの点数が 0～10 点以下の差でばらついている場合，相関係数はいずれも 0.3 未満であり，リツイート数とツイートの面白さに相関は見られない．しかし 0～100 点以上の差でばらつく場合，相関係数は 0.3 以上となり若干だが正の相関が見られる．

4.1 節で実際に得られたツイートの両対数グラフの傾きは 1 未満（べき指数 2 未満）でツイートの価値ではなくリツイート回数に基づいて行動を取ると 6.2 節で述べた．

モデル上でもべき指数が 2 未満である場合,リツイート数と内容の点数には相関は見られなかった．

表 6.1　リツイート数とツイートのおもしろさ（点数）の相関

	ツイート点数 0-1	ツイート点数 0-5	ツイート点数 0-10	ツイート点数 0-100	ツイート点数 0-1000
リツイート頻度 3	0.012 (-0.698)	0.116 (-1.55)	0.230 (-2.159)	0.386 (-3.18)	0.405 (-3.36)
リツイート頻度 5	0.008	0.058 (-1.24)	0.154 (-1.79)	0.388 (-2.98)	0.418 (-3.16)
リツイート頻度 10	0.005 (-0.47)	0.020 (-0.92)	0.058 (-1.26)	0.365 (-3.09)	0.426 (-2.97)

6.4.2　リツイート数とツイート投稿順の相関

図 6.6～図 6.9 はツイート投稿順を横軸,リツイート回数を縦軸に，ツイートの点数を変化させプロットした結果である．

投稿されるツイートの点数の差が大きくなると後から投稿されたツイートにもある程度リツイートされることが示せた．反対にツイートの得点に差がない場合，ツイートの投稿順がリツイート数に大きく影響することが判明した．

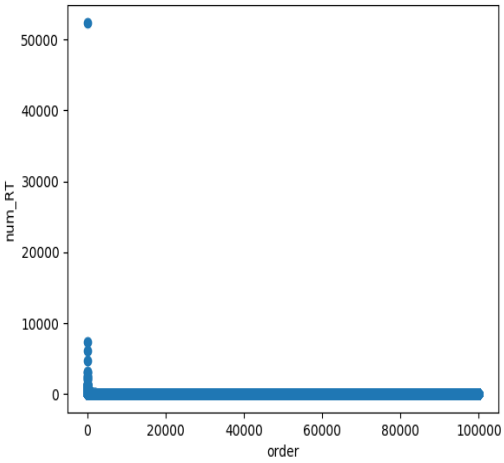

図 6.6：ツイート投稿順とリツイート数の関係
（ツイート点数 0～1）

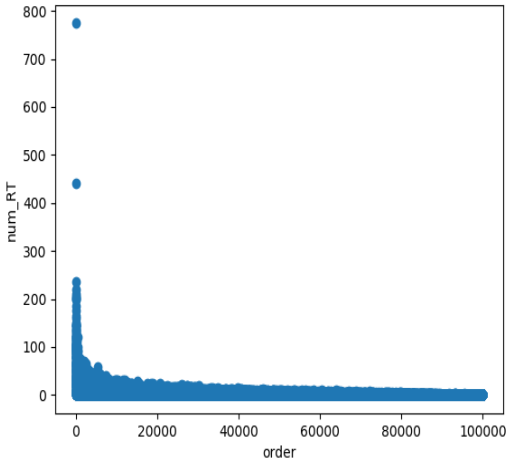

図 6.7：ツイート投稿順とリツイート数の関係
（ツイート点数 0～10）

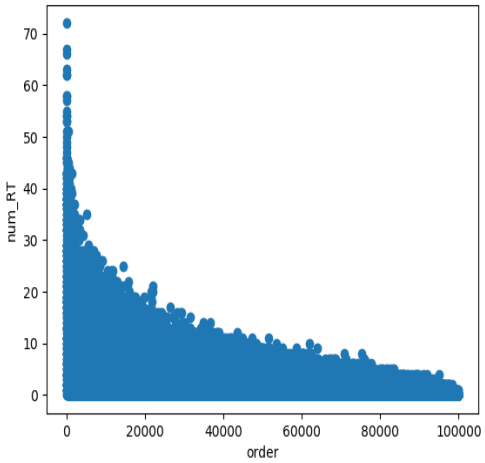

図 6.8：ツイート投稿順とリツイート数の関係
（ツイート点数 0～100）

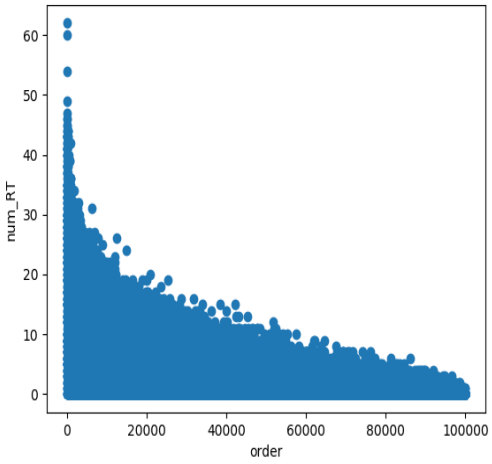

図 6.9：ツイート投稿順とリツイート数の関係
（ツイート点数 0～1000）

第 7 章　結論

本章では各章の総括を行う．また研究を通して発見した新たな課題や方針を示す．

本研究では Twitter API を用いて数々のキーワードでツイートデータを大量に取得，分析した．

災害時や社会的関心の高い情報については得られたツイートの大半がリツイートであり，数少の情報が多数のユーザーに拡散されていることを確認した，
この事実は災害やプロモーションなど，タイムリーかつ広範囲に情報拡散が必要とされる場面で Twitter は有効なツールであるという一例である．

3 章では 1 分間あたりのツイート数を時系列処理して図示した．その結果，イベントに関するツイートではツイート数は昼夜変動とともに増減し，ピークを迎えた後減衰していく傾向が見られた．また，日常,個人的な内容のツイートは昼夜変動でのツイート数の増減は見られたが，日時によるツイート数の増減はなくピークが現れない推移をみせた．

4 章ではキーワード検索により得られたツイートのリツイート数の分布に着目すると，裾の長い分布形状が見られた．これはいかなるツイートにも共通して見られる特徴であり，何らかの共通した要因があることを示す．また，リツイート数と相関のあるメタデータを特定した．5 章では実際の情報拡散過程について昼夜変動を考慮して再現するモデルを提案し，その再現性を示した．6 章では人々のリツイート行動にはリツイート数の大小が影響する，つまり人はリツイート数に重点を置きリツイートする対象を選定するという仮定を置いたシミュレーションモデルを提案し，実際の現象を精度よく再現した．

課題と今後の展望

本研究で用いた確立モデルは単純であり，全ての現実の現象を正確に再現するとはいえない．また，6 章で再現したモデルはネットワークを簡単のため Twitter トポロジーではないものを使用した．今後より正確な再現をするためにも実際のネットワークトポロジーを使用することが望まれる．

Twitter 上でのリツイート数の時間変化は捉えることができたが，リツイートがどのユーザーを経由して拡散されたかという経路の特定は現段階では困難である．経路を特定することで，情報拡散経路の可視化を実現できると考える．

謝辞

本研究を進めるにあたり，ご多忙の中ご指導頂きました塩田茂雄教授に深く感謝申し上げます．また，ゼミを通じて数々の助言をくださった研究室の皆さんには大変お世話になりました．ありがとうございました．

参考文献

[1] 総務省．平成 29 年度版　情報通信白書.
http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h29/html/nc150000.html
2019／01／29 アクセス
[2] 梅島彩奈，宮部真依，荒牧英治，灘本昭代，“災害時 Twitter におけるデマとデマ訂正 RT の傾向”情報処理学会　DBS-152，pp1-6，2011
[3] [Crane, R. and Sornette, “Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System”PNAS，vol.105，pp.15649-15653，2008.
[4] Yang, J. and Leskovec,“Paatterns of Temporal Variation in Online Media”WSDM, pp.177-186，2011
[5] France Cheong and Christopher Cheong. “Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods.2011”.
[6] 石原裕規，諏訪博彦，鳥海不二夫，太田敏澄,“東日本大震災における重要アカウントの抽出とコミュニケーション形態の変容”電子情報通信学会論文誌 Vol.99,No.5,pp.501-513,2016
[7] [Herda Gdelen, Amac, Zuo Wenyun, Gard-MurrayAlexander, Bar-Yam Yaneer, “An exploration of social identity: The geography and politics of news-sharing communities in twitter”, COMPLEXITY,volume19,Issue2,November/December 2013,Pages 10-20
[8] 秦恭史，諏訪博彦，岸本康成，藤原靖宏，新井淳也，飯田恭弘，岩村相哲，鳥海不二夫 ，安本慶一，“クラスタリングに基づく東日本大震災前後の情報拡散度の比較”．人工知能学会
[9] 風間一洋，鳥海不二夫，篠田考祐，“名詞出現頻度の時間的変化に着目した東日本大震災の twitter のトピック分析”，In WebDB forum,2011.
[10] 風間一洋，鳥海不二夫，榊剛史，篠田考祐，栗原聡，野田五十樹，“東日本大震災時の twitter データを用いた単語間の関係の時系列変化の分析”，第 26 回人工知能学会全国大会予稿集 2012.
[11] 三浦麻子　“東日本大震災とオンラインコミュニケーションの社会心理学―そのときツイッターでは何が起こったか―”　電子情報通信学会，95（3），pp.219-223
[12]　三浦麻子，小森政嗣，松村真宏，前田和甫　“東日本大震災時のネガティブ感情反応表出―大規模データによる検討―”心理学研究, *86*(2), 102-111.doi:10.4992/jjpsy.86.13076
[13]　Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011)“ Bad news travel fast: A

content-based analysis of interestingness on Twitter”. WebSci '11 Proceedings of the 3rd International Web Science Conference, Article No. 8. doi:10.1145/2527031.2527052

[14] 安田雪 “ソーシャルメディア上の情報拡散の特性—東日本大震 災時のデマの事例とハブの役割” 関西大学社会学部紀要, 45(1): 33-46 2013

[15] 高野知佐，会田雅樹，“Scaled Laplacian 行列に基づいた固有ベクトル中心性の考察，”電子情報通信学会複雑コミュニケーションサイエンス研究会，CCS2017-12，pp.13-18，2017.

[16] 久保尊広，高野知佐，会田雅樹，“縮退した振動モードから生じる新しいネット炎上モデル，”電子情報通信学会ネットワークシステム研究会，NS2017-80，pp.55-60，2017

[17] “Stanford large network dataset collection,” http://snap.stanford.edu/data/

[18] A.-L. Barabási and R.Albert，“Emergence of scaling in random networks，”Science,vol.286，pp.509-512，1999.

研究成果

1. 南川雅人, 中島圭佑, 塩田茂雄, ″リツイート数分布のべき則性の出現メカニズム,″ 電子情報通信学会 総合大会, 2019 年 3 月（東京）
2. 南川雅人, 中島圭佑, 塩田茂雄, ″Twitter データの特徴分析と人間の行動モデル,″ 電子情報通信学会 ネットワークシステム研究会, 2019 年 3 月（沖縄）
3. 塩田茂雄, 南川雅人, 中島圭佑, “キーワード検索で収集される Twitter データの特徴と Twitter 上での情報拡散過程” 電子情報通信学会 情報ネットワーク研究会, IN2018-64, pp. 31-36, 2018 年 12 月（広島）
4. 南川雅人, 中島圭佑, 塩田茂雄, ″SNS 上の情報拡散過程と投稿数の時間推移のモデル化,″ 第一回 QoS に関する学生技術交流会（名大, NII, 早稲田大, 電通大, 芝浦工大, 千葉大合同）, 2018 年 10 月（東京）
5. 塩田茂雄, 南川雅人, 中島圭佑, ″SNS 上の情報拡散過程と投稿数の時間推移のモデル化,″ 電子情報通信学会 ソサイエティ大会, B-7-38, 2018 年 9 月（金沢）
6. 塩田茂雄, 南川雅人, 中島圭佑, “SNS 投稿件数推移分析のための情報拡散モデルと強相関近似” 第二回計算社会科学ワークショップ, 2018 年 3 月（東京）
7. 南川雅人, 中島圭佑, 塩田茂雄, ″非反応状態を考慮した情報拡散モデルによる SNS 投稿数推移過程の解析,″ 電子情報通信学会 第 5 回コミュニケーションクオリティ (CQ)基礎講座ワークショップ, 2018 年 1 月 20 日（東京）.
8. 南川雅人, 中島圭佑, 塩田茂雄, ″強相関近似による複雑ネットワーク上の情報拡散過程の解析,″ 2017 年度待ち行列シンポジウム「確率モデルとその応用」, pp. 194-195, 2018 年 1 月 19 日（大阪）
9. 中島圭佑, 南川雅人, 塩田茂雄, “SNS における投稿件数推移分析のための情報拡散モデル” 電子情報通信学会 コミニュケーションクオリティ研究会, CQ2017-84, 2017 年 11 月 17 日（高松）
10. 南川雅人, 塩田茂雄, ″複雑ネットワーク上の情報拡散過程におけるノード相関の影響,″ 電子情報通信学会 ソサイエティ大会, B-7-3, 2017 年 9 月 12 日（東京）
11. 南川雅人, 塩田茂雄, “ネットワーク上の情報拡散過程におけるノード相関の影響” 電子情報通信学会 コミニュケーションクオリティ研究会, CQ2017-58, pp. 43-58, 2017 年 8 月 29 日（東京）
12. 南川雅人, 塩田茂雄, ″ネットワーク上の情報拡散モデルにおける因果律の破れ,″ 電子情報通信学会 コミニュケーションクオリティ学生ワークショップ, 2017 年 8 月 26 日（日本工業大学: 埼玉）