Twitter 上の人間関係ネットワークの抽出とその分析

平成 23 年度卒業論文

千葉大学都市環境システム学科
指導教官：塩田茂雄

08T0230F
晒谷亮輔

目次

第 1 章　序論

1.1　研究の目的と背景

私達が情報を得るときには身近な人からの情報に大きく影響を受ける. その情報の取得手段はインターネットの普及と共に多様化している. 近年急速に普及しているネット上の様々な SNS（ソーシャルネットワーキングサービス）は，今や情報共有やコミュニケーションインフラとして必要不可欠なものとなっている．その中でも twitter は日本国内でも本格的な流行の様相を見せており，基本的に招待加入性である従来の SNS とは異なった，自らの興味関心に基づき情報を取得，発信することができる新たな情報源として注目を集めている. その利用者数は日本では 1700 万人, 全世界では 2 億 1000 万人の大台を超えており，1 日あたりの平均ツイート数は 2 億を超えている．純粋なユーザー以外にも twitter と連携したウェブサービス，プロモーション活動などネット業界全体やビジネスシーンにおいても twitter は大きな存在感を見せている．また，震災時にソーシャルメディアが電話に代わるコミュニケーションインフラとして活用されるなど，災害における強さも再認識された.

そのような中, 2007 年に Balachander Krishunamurthy　ら[1][2][3]は初めて本格的な twitter のネットワーク特性の分析を行った．Twitter のネットワーク分析で最も代表的なものは Haewoon Kwak ら[4]が 2009 年の取得データをもとに行った分析であり, twitter ネットワークのスケールフリー性，スモールワールド性を証明した．それ以降 twitter のネットワーク特性に強く着目した分析は行われていない．2007 年の twitter ユーザー数は世界で 50 万人，1 日あたりの平均ツイート数は 5000，2009 年の twitter 利用者数は 2000 万人，1 日当たりの平均ツイート数は 250 万であり，現在のユーザー数と見比べると, 最新の研究が行われた 2009 年から 2011 年にかけて twitter のネットワークは急激に成長している．また，日本のユーザー数に関しては，4 倍近く増加しており，本研究では Twitter のネットワークが拡大し利用方法も多様化している今，改めて twitter ネットワークの抽出を行い，twitter のネットワークの分析を行う．データ収集では日本ユーザーを多く含めたデータから twitter のフォロー，フォロワーでのつながりに着目し twitter 上の新たな人間関係ネットワークの抽出からソーシャルグラフの解析を行っていく.

1.2　論文構成

本論文では以下のような構成をとる．第 1 章では序論として研究背景，目的の解説，第 2 章では研究対象である twitter の持つ特徴，基本事項の解説，第 3 章では twitter 上の人間関係ネットワークの抽出方法の解説，第 4 章では次数分布，相関関係，クラスター分析から，twitter 上のネットワーク特性の分析を行った．第 5 章では分析結果のまとめ，結論を述べる．

第 2 章　Twitter

2.1　概要

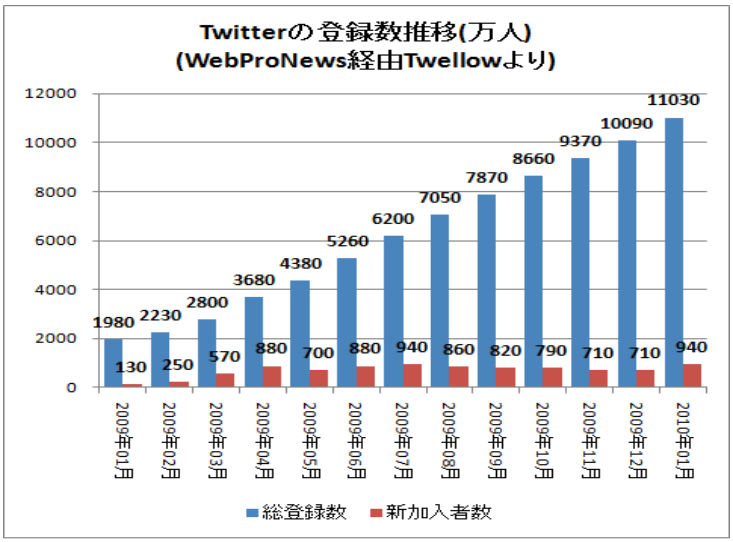
twitter[1]はリアルタイムの情報ネットワークで, 140 文字以内の「ツイート」(tweet) と称される短文を投稿できるマイクロブログサービスである．Twitter の心臓部はツイートであり，文字以外にもそのツイート内にリンクを挿入することで写真や動画，その他のメディアコンテンツを共有することができる．

従来の SNS である mixi は友人として登録されなければ，制限されたプロフィール内容以外の情報は公開さらないが，twitter のツイート内容はパブリックに公開され，twitter に登録すると全世界のパブリックユーザーのツイート情報を見ることができる．

Twitterのユーザー数は2011年12月時点で全世界で2億1000万人, 日本国内で1800万人（推定値）である．図 2.1 に twitter の世界ユーザー推移（2009/01～2010/01）のグラフを，図 2.2 に twitter 国内ユーザー数推移のグラフを示す．尚ユーザー数とは登録者数ではなく twitter.com に訪問した数である訪問者数のことを指す．また，図 2.2 グラフ上の数値はモバイル機器を除いた PC から twitter 公式サイトにログインしたユーザー数である．また，世界から見た日本の twitter 利用状況を量る指標として図 2.3 に世界の twitter リーチ率ランキングを記す．

図 2.1 より 2010 年 01 月のユーザー数から現在のユーザー数 2 億人と比べるとこの 1 年でユーザー数はさらに倍となっている．また，図 2.2 より，国内の SNS では一番多くのユーザー数を獲得していることが分かる．さらに世界の人口に対する twitter のアクセス数を表したリーチ率では日本は世界で 4 位と，日本は twitter の中心利用国となっている．

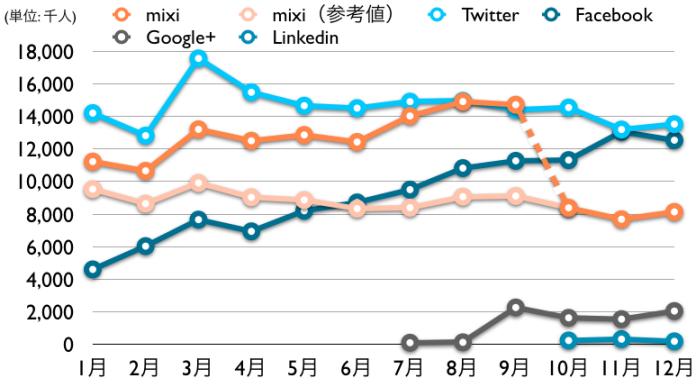
また，企業ソーシャルメディア利用動向調査報告書 2011 によると，何らかのソーシャルメディアに取り組んでいる企業において、取り組んでいるソーシャルメディアは「Twitter」が 58.0%でトップである．以下、「ブログ」44.5%、「Facebook」38.7%、「mixi」（35.0%）と続く。このように，Twitter は企業のプロモーション活動やマーケティングなどのビジネスシーンにおいても利用されている．



↑ *Twitterの登録数推移(万人)(WebProNews経由Twellowより)*

図 2.1：twitter 世界ユーザー数推移（2009/01～2010/01）

■ PC訪問者数推移



(単位: 千人)

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
mixi	11,228	10,659	13,211	12,507	12,864	12,433	14,033	14,917	14,723	8,385	7,684	8,135
Twitter	14,211	12,824	17,571	15,489	14,666	14,516	14,914	14,962	14,416	14,551	13,199	13,529
Facebook	4,598	6,030	7,659	6,939	8,204	8,717	9,504	10,827	11,274	11,319	13,061	12,543
Google+							91	166	2,257	1,622	1,541	2,038
Linkedin										228	310	176

図 2.2：twitter 国内ユーザー数推移（2011/01～2011/12）

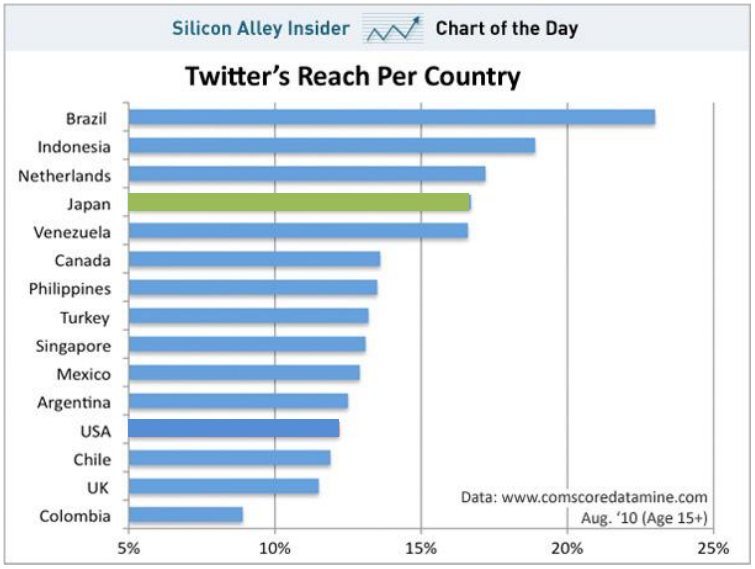


図 2.3：世界の twitter リーチ率ランキング

2.2　用語

2.2.1　ユーザー／スクリーン名（user / screen name）

Twitter の利用者，またはそのアカウントをユーザーと呼ぶ．ユーザーは@で始まるスクリーン名を設定し区別する．スクリーン名は最大 15 文字のアルファベット（大文字小文字），数時，アンダーバーで構成される．

2.2.2　ツイート（tweet）

Twitter における「つぶやき」を表す．1 つのツイートには 140 字という制限が設けられている．

2.2.3　タイムライン（timeline）

リアルタイムの時系列によって表示されるツイートの一覧である．Twitter にログイン後表示されるホームタイムラインには自らのツイートとフォローしているユーザーからのすべてのツイートが表示される．

2.2.4　フォロー（follow）

特定のユーザーのツイートを自分のタイムライン上で閲覧できるよう登録することである．

2.2.5　フォロワー（follower）

特定のユーザーをフォローしているユーザーのことを指す．ユーザーA がユーザーB をフォローしている場合，ユーザーA はユーザーB のフォロワーである．

2.2.6　フレンド（friend）

フォロワーと反対に，特定のユーザーがフォローしているユーザーのことを指す．ユーザーA がユーザーB をフォローしている場合，ユーザーB はユーザーA のフレンドである．

フォロー，フォロワー，フレンド，タイムラインの関係を図 2.4 に記す．この場合ユーザーB をフォローしているユーザーA（フォロワー）のタイムラインにユーザーB のツイートが表示される．

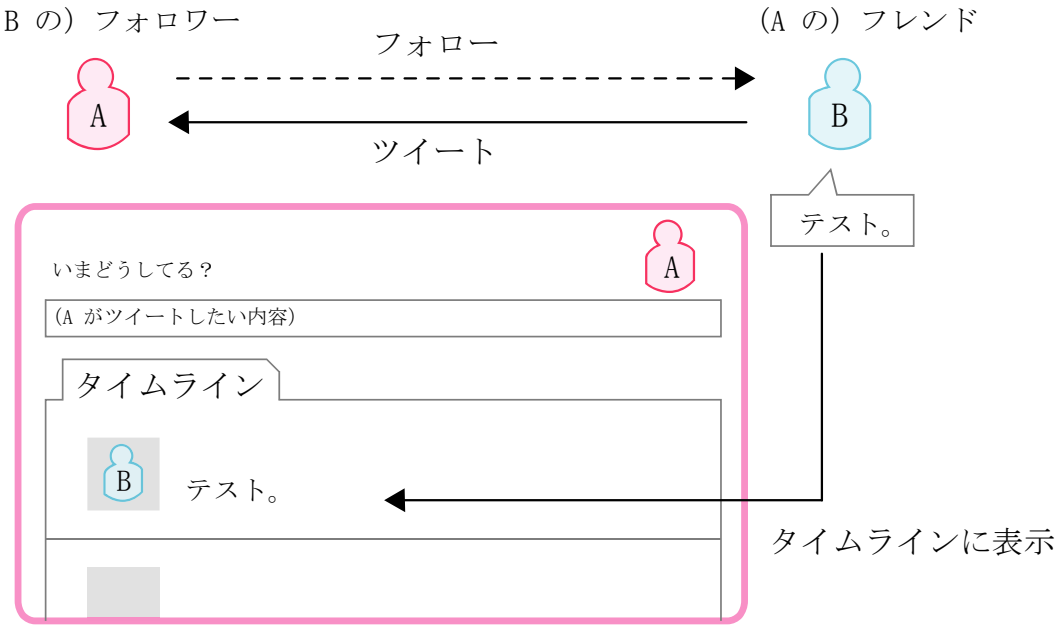


図 2.4　フォロー，フォロワー，フレンド，タイムラインの関係

2.2.7　関連ツイート（mention），リプライ（reply）

特定のユーザーに宛てたツイートや，特定のユーザーについてツイート内で触れたい場合に「@スクリーン名」を含めてツイートされたものを関連ツイートと呼ぶ．自分宛ての関連ツイートは，ツイートしたユーザーをフォローしていなくても「@関連」として一覧表示される．一方，フォローしているユーザーのツイートであっても，「@スクリーン名」がツイート本文の最初に現れる関連ツイート（リプライとも呼ばれる）は，その対象アカウントをフォローしていないユーザーのタイムラインには表示されない．

関連ツイートのしくみを図 2.5 に記す．ユーザーA から B 宛ての関連ツイートは，ユーザーB の@関連に一覧表示される．ユーザーA，B をフォローしているユーザーD のタイムラインにも同様に表示される．しかし，ユーザーA だけをフォローしていて，ユーザーB をフォローしていないユーザーC のタイムラインには関連ツイートは表示されない．

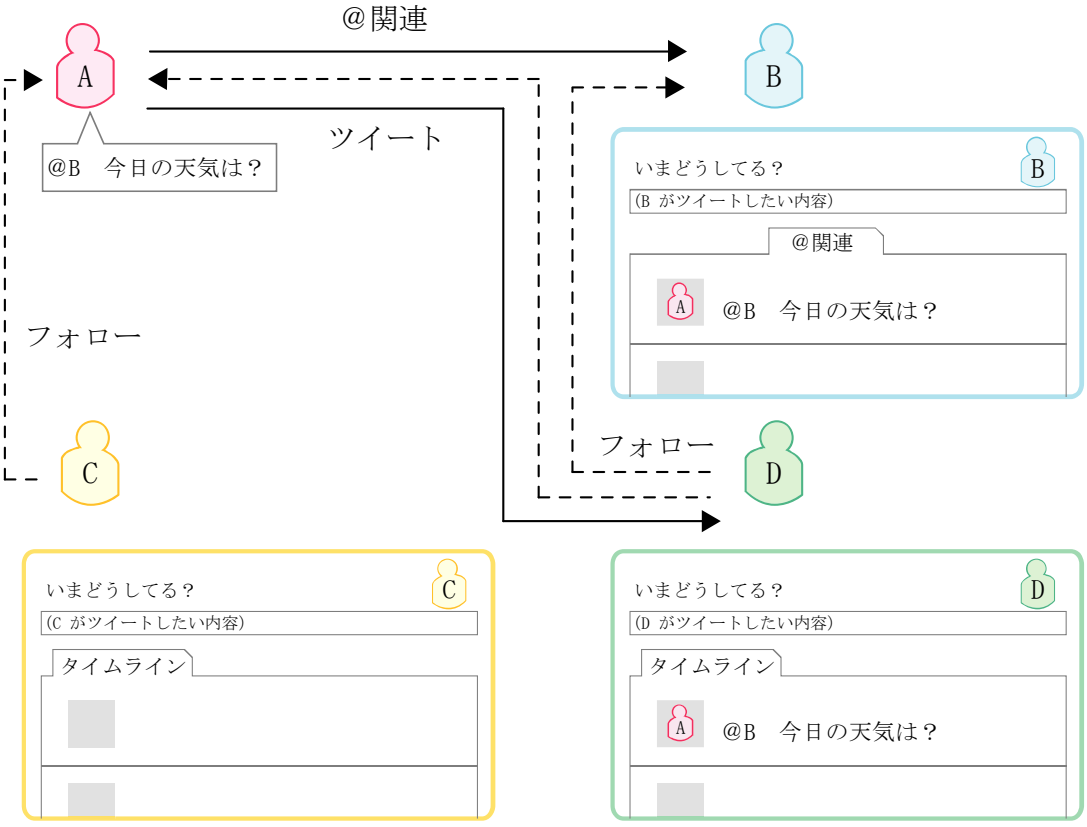


図 2.5　関連ツイートの仕組み

2.2.8　リツイート（retweet）

他ユーザーのツイートを自分のフォロワーに向けて再発信する操作，また再発信されたツイートのことを指す言葉である．またリツイートには公式リツイートと非公式リツイートの 2 種類存在する．

公式リツイートとはオリジナルのツイートの内容を変えずにそのまま再発信するリツイートであり，非公式リツイートとは「RT@スクリーン名：＜元のツイート＞」というようにオリジナルのツイートに対してその投稿者名と自分のコメントを追加して再発信するリツイートである．

公式ツイートは関連ツイートとして扱われないため，オリジナルのツイートをしたユーザーのタイムライン上には表示されないが，非公式リツイートはオリジナルのツイートを行ったユーザーのタイムラインに表示される．

公式リツイートのしくみを図 2.6 に，非公式リツイートのしくみを図 2.2.3 に記す．

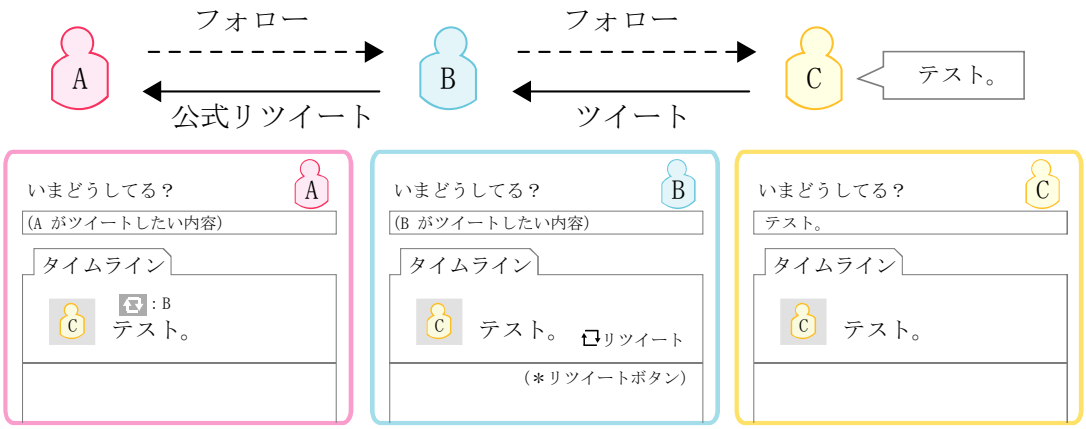


図 2.3：公式リツイート

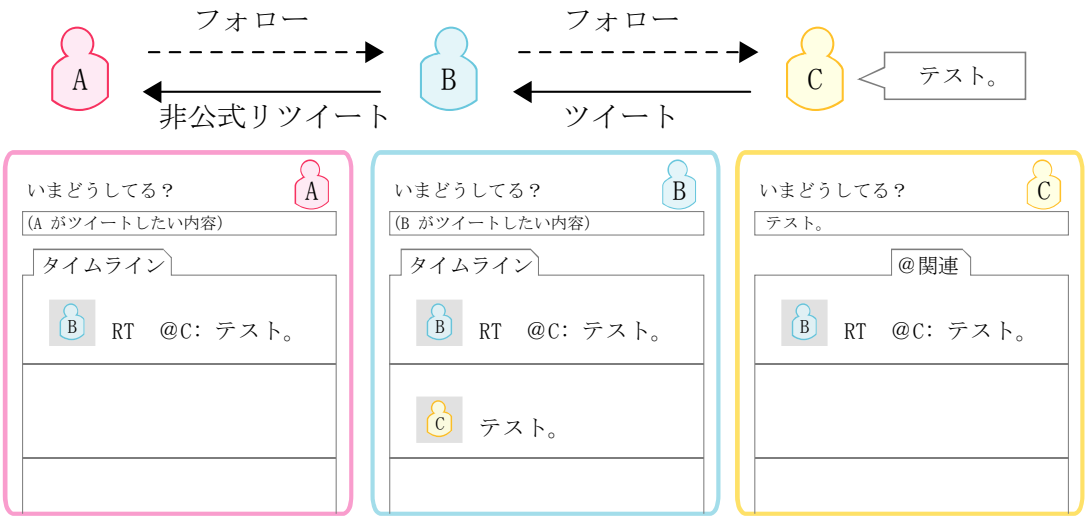


図 2.6：非公式リツイート

2.2.9　リスト（list）

フレンドとは別にユーザーをグループ化し，まとめて管理する機能である．1 つのリストの制限は 500 人までとなっており，各々が作成したリストを参照することでそのリスト内のフレンドのみのタイムラインを表示することが可能となる．また，作成したリストは他ユーザーに公開され，他ユーザーのリストを購読することも可能である．

2.2.10　ハッシュタグ（hash tag）

ハッシュタグとはツイートに含める＃から始まる文字列であり，特定のイベントやトピックに対してツイートする際に検索対象としてハッシュタグを用いる．ハッシュタグをつけているツイートはそのハッシュタグによりカテゴライズされ，そのハッシュタグを検索すると，同一のイベントに参加しているユーザーや同一の興味関心を持つユーザーのツイートが時系列として一覧表示される．

2.2.11　ツイートプライバシー（tweet privacy）

ツイートを非公開にする設定を指す言葉である．この設定を行うことで，許可したフォロワーのタイムラインにのみ自身のツイートを公開することが可能となる．非公開設定のユーザーをフォローする場合は相手に対しフォローリクエストを送信し許可を得ることが必要となる．

第３章　ネットワークの抽出

3.1　Twitter API

3.1.1　概要

API とは Application Programming Interface の略で，サービスやソフトウェアをプログラムから利用する為の命令である．Twitter は，Web 技術を使ったサービスを，ブラウザからだけでなく，Web を通してプログラムからも利用できるようにした WebAPI を公開することで，Twitter と連携する多数のデスクトップモバイルアプリケーションや Web サイトの開発を促している．

TwitterAPI [2]は大きく REST API，検索 API，ストリーミング API，Web サイト向け API の 4 種類に分けられる．本研究では主にタイムラインやソーシャルグラフなどを取得する参照系メソッドがある REST API を利用し twitter のユーザー情報を取得することでネットワークの抽出を行った．

3.1.2　API 制限

Twitter API を利用する場合，ユーザーの IP アドレス単位で API 実行回数制限が設けられている．その制限回数は通常 60 分間に 150 回までであり，OAuth 経由でアクセスした場合 60 分間に 350 回までとなる．また Twitter の運用状況によってその回数制限は変動する．

3.1.3　OAuth 認証

OAuth は，The OAuth Community[3]が策定する認証方式である．現在 Web サービスで標準となりつつある認証方式で，Flickr や Twitter，Google などが採用している．Twitter が採用している OAuth の方式は 3-legged OAuth[4]と呼ばれるものである．

OAuth はユーザーがアプリケーションにパスワードを渡すことなくアカウントへのアクセスの可否を決定できる安全な認証方式であり，API 実行回数制限が 350 回/h と通常よりも多く API を実行することが可能であるため本研究では OAuth 認証を利用した．

3.2　データ取得

3.3.1　取得データ

twitter 上の人間関係ネットワークの抽出を行うために，実際のツイッター上のユーザー情報の内ユーザーID，フォロー数，フォロワー数，ツイート数，フォロワー一覧 ID の 5 種類の項目に対してデータ収集を行った．

ユーザーID，フォロー数，フォロワー数，ツイート数に関しては自らのツイッターアカウントである「@sarashi_ya」を始点に各フォロワーをたどり 66,665 ユーザーのデータを取得した．これをデータセット A とする．また，分析に偏りがでないようにランダムに指定したユーザーからもそれぞれ 56,736 ユーザーのフォロー数，フォロワー数，ツイート数のデータを取得した．これをデータセット B とする．

また，ツイートユーザー相互の関係性（クラスタ係数等）を分析するためのデータとして，「@sarashi_ya」を始点にフォロワーをたどって取得した 10,000 を重複カットした 7,713 ユーザーそれぞれのフォロワー一覧 ID（47,139,638 個）を取得した．これをデータセット C とする．このデータセット C を用いて 7,713 ユーザー間のフォロー，フォロワー関係を調査した．

3.2.2　取得方法

データの取得には前項で述べた Twitter API の REST API を利用しデータを取得した．データ収集プログラムは PHP と JAVA の 2 種類でそれぞれ構成し tmhOAuth，Twitter4j の 2 つのオープンソースライブラリを利用した．使用した API メソッドとその効果に関して表 3.1 にまとめる．

表 3.1：API メソッドとその効果

メソッド	効果
fllowers/ids (http://api.twitter.com/1/fllowers/ids.format)	指定したユーザーのフォロワー一覧 ID を取得する．
statuses/followers (http://api.twitter.com/1/statuses/followers/ids.format)	指定したユーザーのフォロワーの最新ステータスを取得する．

第 4 章　ネットワークの分析

4.1 次数分布

4.1.1　定義

点（ノード）と線（リンク）からなるグラフを考える．ノードの張るリンク数を次数と呼ぶ．（ノードnがk本のリンクを張るとき，ノードnの次数はkであるという．

次数分布[2]とはノードの次数（リンク数）の分布であり，次数kを持つノードの割合（ノードの次数がkに等しい確率）$p(k)$で定義される．次数分布がベキ則に従う，つまり

$$p(k) \propto k^{-\eta} \tag{1}$$

を満たす場合，ネットワークはスケールフリー性を持つと呼ばれる．ηをベキ指数と呼ぶ．スケールフリー性を持ったネットワークはスケールフリーネットワークと呼ばれ，現実世界には数多くのスケールフリーネットワークが存在する．スケールフリー性を持つネットワークの特徴は，一部のノードが他の多数のノードとエッジで繋がっており，大きな次数を持っている一方で，大多数のノードはごくわずかなノードとしか繋がっておらず次数は小さく少ないリンクしか持っていないということである．

実世界のスケールフリーネットワークのベキ指数の例を表 4.1 に挙げる．

表 4.1：実世界のスケールフリーネットワークのベキ指数

ネットワーク	ベキ指数
WWW（ワールド・ワイド・ウェブ）	1.9～2.7
インターネット	2.1～2.5
映画俳優の共演ネットワーク	2.3～3.1
性的関係のネットワーク	3.2～3.4
タンパク質の反応ネットワーク	2.4～2.5

4.1.2　フォロー数，フォロワー数分布

Twitter 上に形成される人間関係ネットワークは有向グラフであり，フォロー数は出次数，フォロワー数は入次数に相当する．従って，次数分布もそれぞれに定義される．そこで，データセット A およびデータセット B のフォロー数（出次数），フォロワー数（入次数）の次数分布から，フォロー数，フォロワー数のそれぞれにスケールフリー性が存在するかを分析する．

表 4.2 にデータセット A および B のフォロー数，フォロワー数，最大値，最小値，平均値をまとめた．

表 4.2：フォロー数（出次数），フォロワー数（入り次数）のデータ情報

		データセット A	データセット B
データ総数（人）	重複カット前	106796	68385
	重複カット後	66665	56736
フォロー数（人）	最小値	1	1
	最大値	739241	738622
	平均値	3738	4416
フォロワー数(人)	最小値	0	0
	最大値	1718299	6241728
	平均値	3799	4017

データ取得方法がフォロワーをたどり取得する方法であったため，フォロー数の最小値は 1 となっている．

データセット B のフォロワー数最大値は「TwitPic」

データセット A のフォロ数，フォロワー数それぞれのユーザー数割合をまとめたグラフを図 4.1 に示す．データセット B に関しても同様に図 4.1.2 に示す．

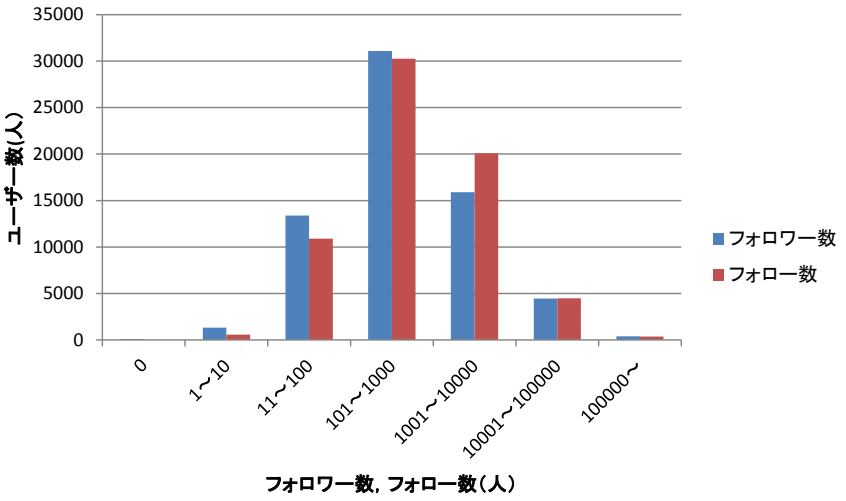


図 4.1：フォロー数，フォロワー数のユーザー数割合（データセット A）

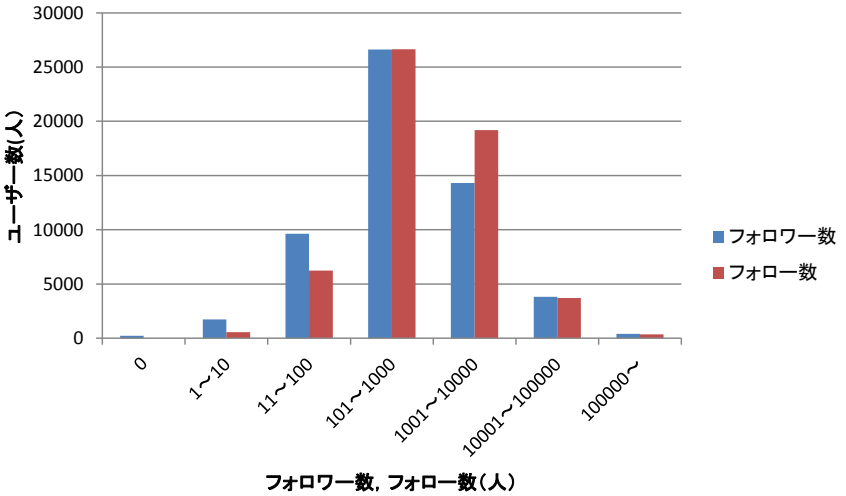


図 4.2：フォロー数，フォロワー数のユーザー数割合（データセット B）

データセット A，データセット B，両グラフ共に全体のフォロー数，フォロワー数の全体の割合として，4.2 からフォロワー数，フォロー数がそれぞれ 101～1000 人の範囲にユーザーが集中していることが分かる．

また，特徴の一つとして両グラフ共に 1001～10000 人の範囲でユーザー数割合がフォロワー数よりもフォロー数のほうが大きいことがあげられる．そこで，1001～10000 人の範囲内のフォロー数，フォロワー数それぞれのユーザー数割合をまとめたグラフを図 4.3（データセット A），図 4.4 に示す．

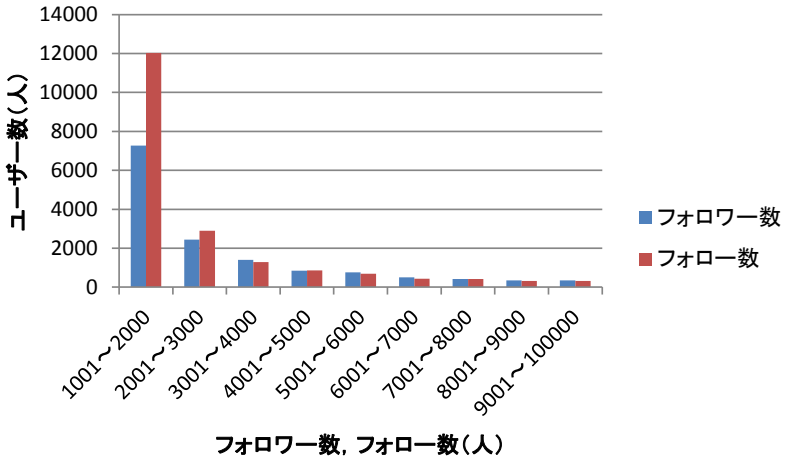


図 4.3：フォロワー数，フォロー数のユーザー数割合（デートセット A）

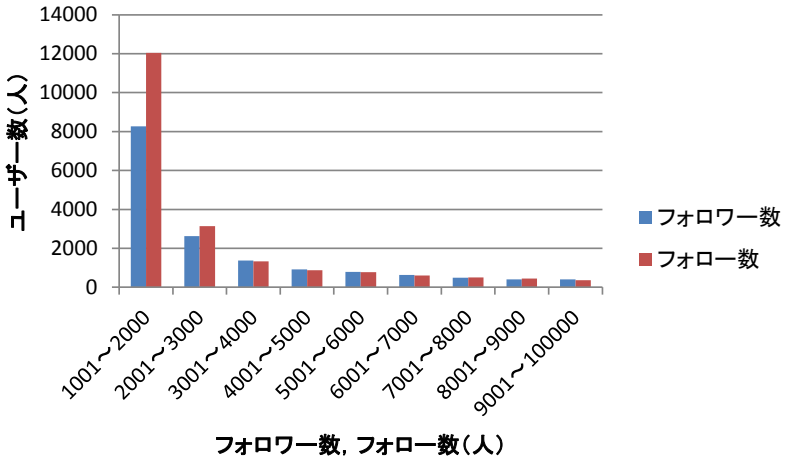


図 4.4：フォロワー数，フォロー数のユーザー数割合（データセット B）

図 4.3, 図 4.4 から 1001～10000 の範囲でユーザー数割合がフォロワー数よりフォロー数のほうが小さかった，その差のほとんどが 1001～2000 人の範囲内で生じていたことが分かる．

その要因となる可能性としてフォロワーが 2000 人以上いない人は 2000 人以上をフォローすることができないという twitter に存在するルールがあげられる．

さて，図 4.5 にデータセット A および B のフォロー数（出次数）の分布を，図 4.6 にデータセット A および B のフォロワー数（入次数）の分布を示す

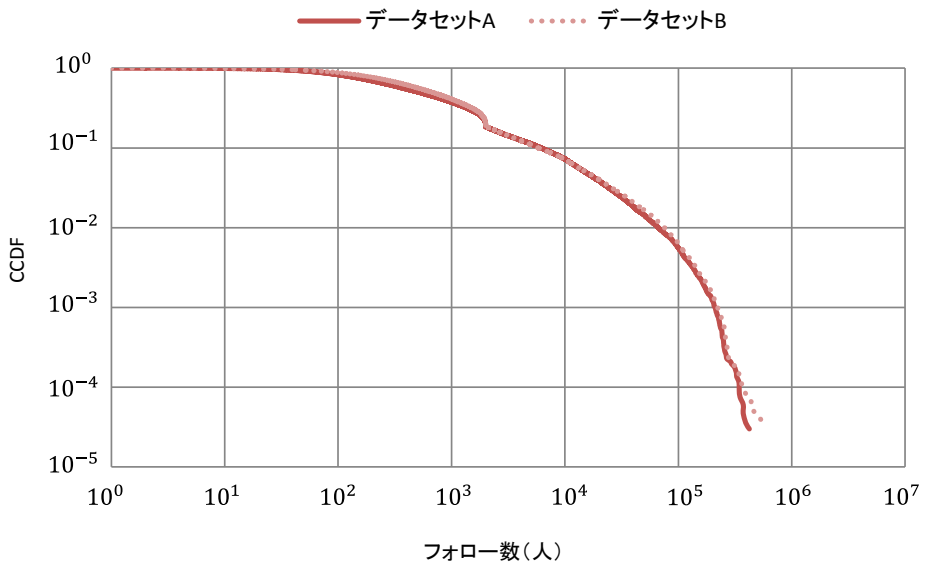


図 4.5：データセット A および B のフォロー数（出次数）の分布

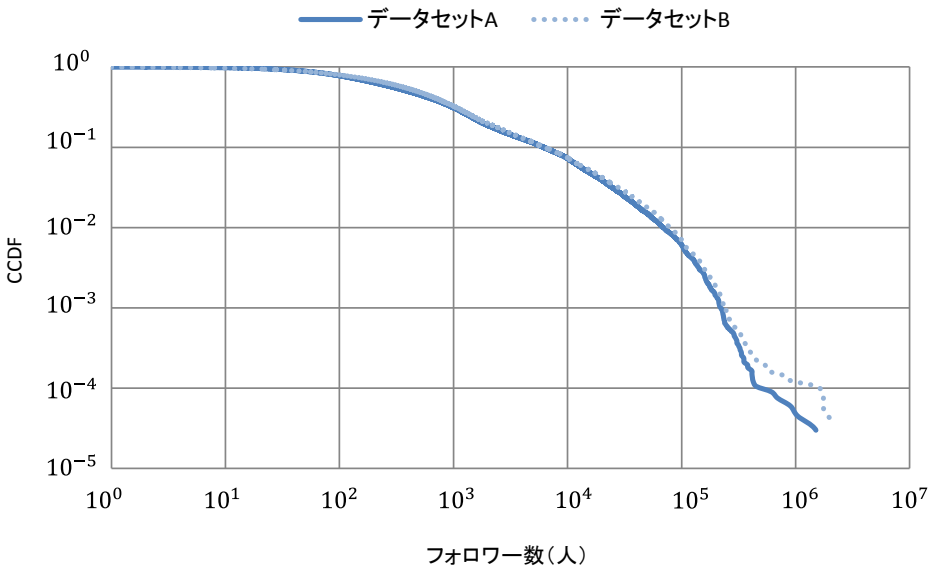


図 4.6：データセット A および B のフォロワー数（入次数）の分布

データセット A とデータセット B ともに，フォロー数，フォロワー数の次数分布の分布は類似したものとなっている．

次にデータセット A のフォロー数，フォロワー数分布の比較図を図 4.7 に，データセット B のフォロー数，フォロワー数分布の比較図を図 4.8 に示す.

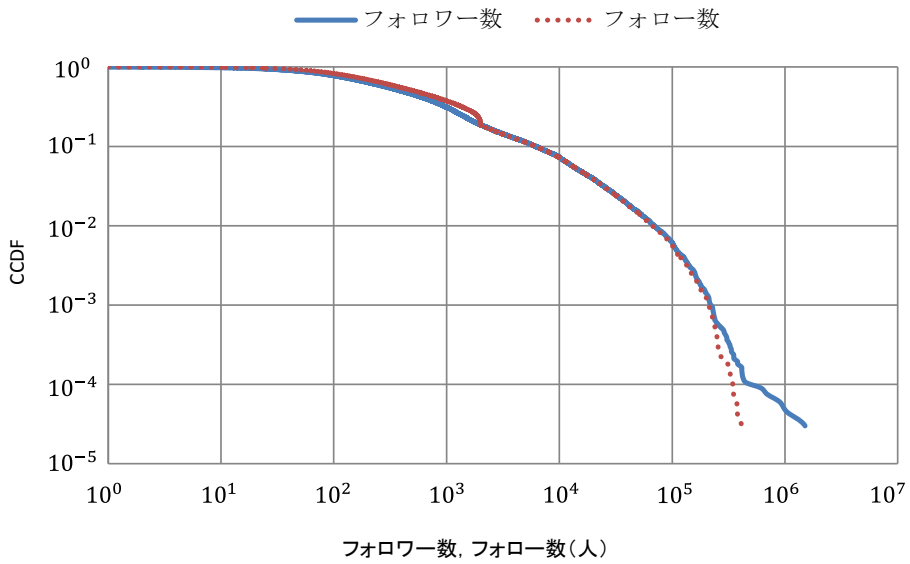


図 4.7：フォロー数，フォロワー数分布の比較図（データセット A）

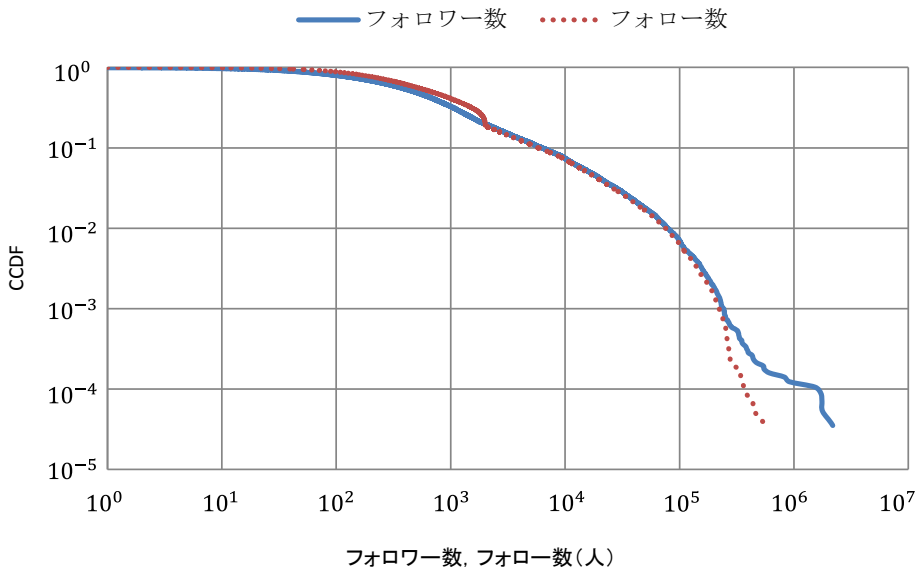


図 4.8：フォロー数，フォロワー数分布の比較図（データセット B）

図 4.7 および 4.8 から分かるようにフォロワー数，フォロー数ともにその　分布は必ずしもベキ側に従っておらず，twitter 上のネットワークがスケールフリーでない可能性を示唆する結果となった.

4.2　ツイート数の分布

ツイート数とはつぶやきである「ツイート」の過去からの総数である．ツイート数はツイッターのアカウントを作成した段階から蓄積されていくが, 各ユーザーが自分のツイートを削除した場合，総数も削除した数だけ減少する．したがって，今回収集したツイート数の全データが完全に過去からのツイートの総数であるとは断定できないことを踏まえて分析を行う. ここではツイート数の分布からツイート数ごとのユーザー数の割合を分析する.

まず，データセット A およびデータセット B のツイート数の最大値，最小値，平均値を表 4.3 に記す.

表 4.3 : データセット A，データセット B のツイート数のデータ情報

		データセット A	データセット B
データ総数（人）	重複カット前	106796	68385
	重複カット後	66665	56736
ツイート数（回）	最小値	0	0
	最大値	729328	749012
	平均値	4521	7349

データセットAのツイート数729328回のユーザーはBOTで,「たろっとさん@tarot3」である．BOT とは，twitter API を利用しプログラムで自動的にツイートされるように設定されたアカウントのことである．例えば「たろっとさん@tarot3」の場合，@tarot3 に関連ツイートを行った場合に自動的に@tarot がタロット占いをのせた関連ツイートを返すようするように設定されたプログラム BOT である．このようにツイッターには数多くの BOT が存在する.

データセット A のツイート数ごとのユーザー数の割合を比べたグラフを図 4.9 に示す．同様に，データセット B のツイート数ごとのユーザー数の割合を比べたグラフを図 4.10 に示す．

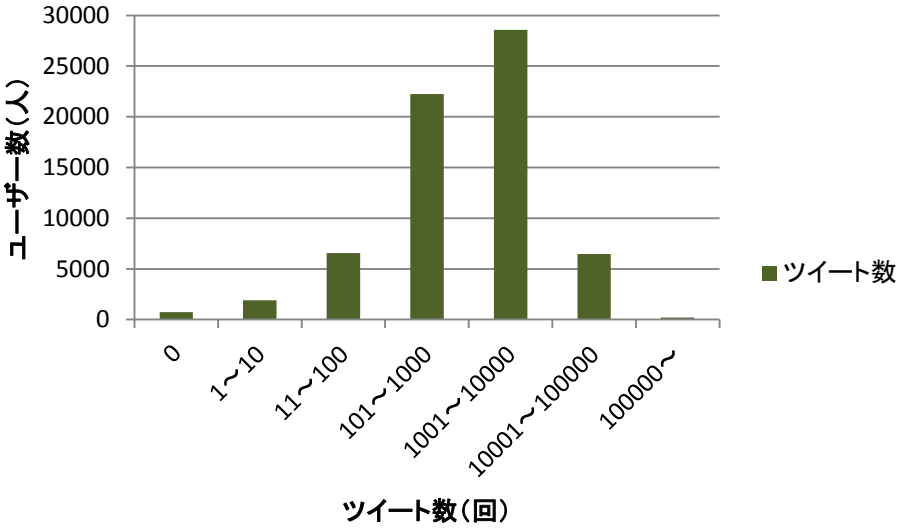


図 4.9：ツイート数のユーザー数割合（データセット A）

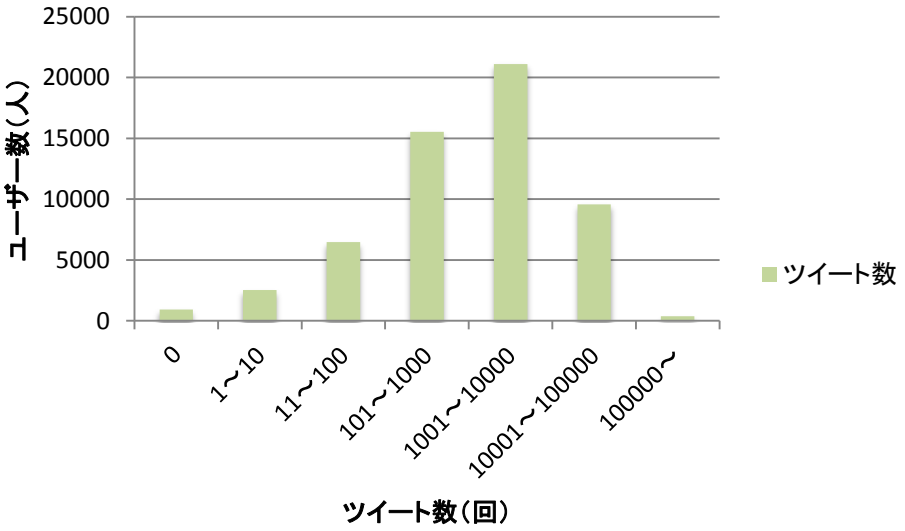


図 4.10：ツイート数のユーザー数割合（データセット B）

データセット A，B 共に，ツイート数が多くなるにつれてそのユーザー数の割合いも高くなっていく．ツイート数 1001～10000 回の範囲にユーザー数が最も集中しており，ツイート数 10000 回を超えるとユーザー数の割合は 2 分の 1 以下に減少する．ツイート数 100000 回以上のユーザーの割合はデータセット A，B それぞれ 0.31%，0.74%である

4.3　フォロワー数，フォロー数，ツイート数の相関関係

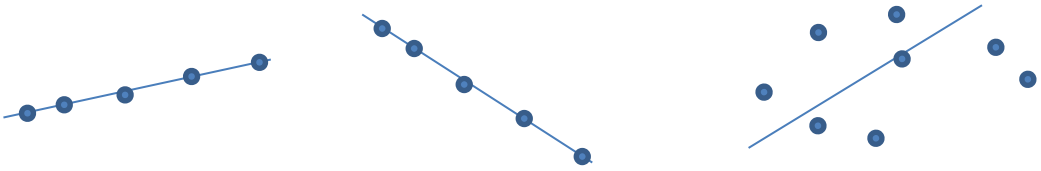
4.3.1　相関係数の定義

相関係数とは，2 つの確率変数の間の相関を示す統計学的指標である．原則，単位はなく，-1 から 1 の間の実数値をとり，1 に近い場合は 2 つの確率変数には正の相関があるといい，-1 に近い場合は負の相関があるという．また，0 に近いときはもとの確率変数の相関は弱い．

データ列$(x_i,\ \mathrm{y_i})(i=1,2,3,\ldots,n)$があたえられたとき相関係数 r は式(2)で定義される．ただし $\bar{x},\bar{y}$ はそれぞれ $X=\{x_i\},Y=\{y_i\}$ の相加平均である．

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2}$$

$r=1$ ならばすべての観測値が正の傾きをもつ同一直線上に並び，$r=-1$ ならば，すべての観測値が負の傾きを持つ同一直線上に並ぶ．図 4.11 に相関係数 r の値による直線関係を示す．また，相関の強さを経験的に判断する基準値を表 4.4 に記す．



$r=1$　　$r=-1$　　$r=0.5$

図 4.11：相関係数による直線関係

表 4.4：相関係数の判断基準

負の相関	相関の強さの判定	正の相関
-1～-0.7	強い相関がある	+1～+0.7
-0.7～-0.4	中程度の相関がある	+0.7～+0.4
-0.4～-0.2	弱い相関がある	+0.4～+0.2
-0.2～0	ほとんど相関がない	+0.2～0

また，r を 2 乗すると「説明率」になり,%として読むことができる.
例えば,相関係数が 0.70 であれば,説明率は 0.702=0.49 であり,一方の変数が他方の変数の 49%の動きを説明することがわかる.（どちら側の変数からいっても同じ）

4.3.2 フォロワー数，フォロー数，ツイート数の相関係数

(2)式に基づき，フォロー数，フォロワー数，およびツイート数相互の相関関係を分析した.

表 4.5 はデータセット A および B におけるフォロー数，フォロワー数，ツイート数相互の相関係数を示したものである.

表 4.5 フォロー数，フォロワー数，ツイート数の相関係数

	データセット A	データセット B
フォロー数ー，フォロワー数	0.8739	0.5259
フォロー数，ーツイート数	0.1167	0.0755
フォロワー数，ーツイート数	0.0987	0.0408

表 4.4 の基準値から見ると，（フォロー数，フォロワー数）および（フォロワー数，ツイート数）の相関係数は 0 に近いことからほとんど相関がなく，互いに独立であることが分かる．この結果はツイートを多数回行うことが必ずしも twitter 上の人間関係の形成に寄与しないことを意味している.

また，フォロー数とフォロワー数の相関係数はデータセット A，B 共に 0.5 を超えており, 明確に正の相関を示す結果となった. これはフォロー数の多い (少ない) ユーザーは自らがフォローする人も多い（少ない）ということを意味している.

4.3.3　フォロワー数，フォロー数，ツイート数の散布図

相関係数により相関関係を判断する上で，いくつか注意すべき点がある．

1 つ目が曲線相関である．相関係数の値が±1 に近いということは X と Y の間に強い直線関係があることを意味している．その値の大小は直線関係の強さを表すものであるが，値が 2 倍であるから，2 倍の相関があるとは言えない．また，0 に近いということは直線関係がないことを意味しているが必ずしも無関係であると断定することはできない．

図 4.12 に示す山形の曲線相関では，相関係数が 0 であるが強い相関関係が存在する．

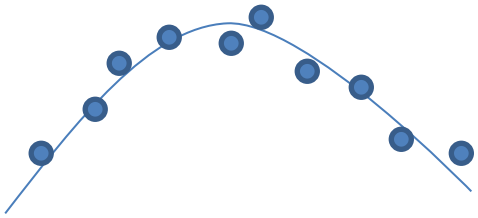


図 4.12：曲線相関

2 つ目に外れ値の存在に注意しなければならない．相関では，他から遥かに離れた点である，外れ値の有無が相関係数に大きな影響を与える．相関係数が 0 に近かったとしても，外れ値の存在を考慮しなければ低い相関であると断定することはできない．

以上のことから，フォロワー数，フォロー数およびツイート数相互の散布図も含め，フォロー数，フォロワー数およびツイート数相互の相関関係を分析する．

データセットAにおけるフォロワー数，ツイート数の散布図を図 4.13 に，データセットBにおけるフォロワー数，ツイート数の散布図を図 4.14 に示し，フォロワー数とツイート数の関係性を分析する．

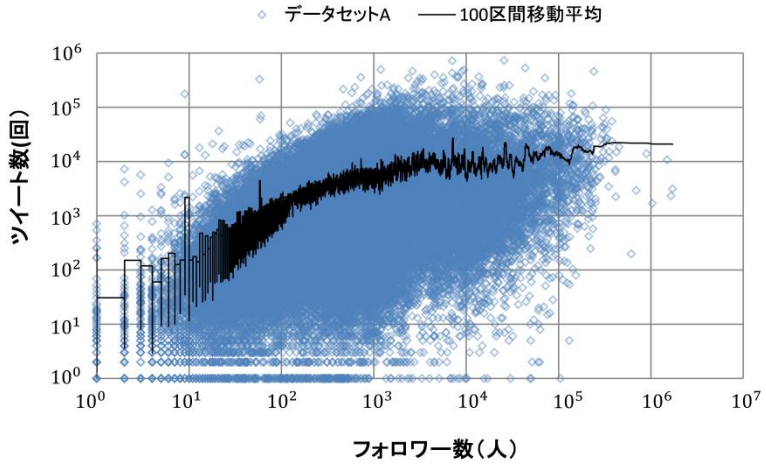


図 4.13：データセットAに対するフォロワー数，ツイート数の散布図

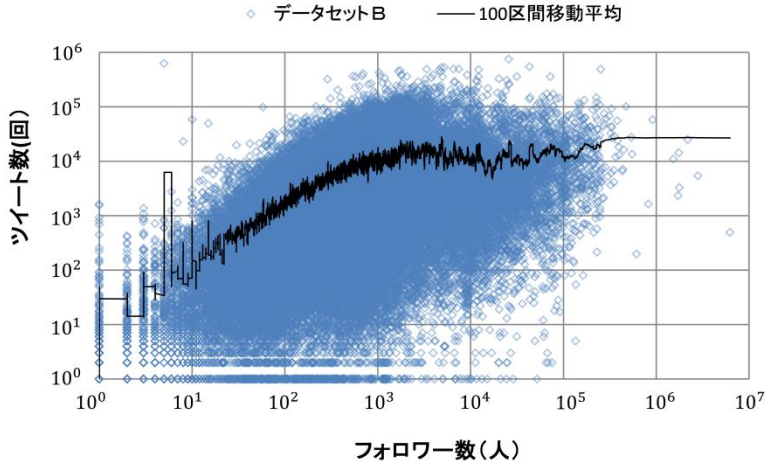


図 4.14：データセットBにおけるフォロワー数，ツイート数の散布図

グラフには 100 区間ごとの移動平均（1 番目から 100 番目までのデータの平均値を区間ごとにだしたもの）をプロットした．

フォロワー数とツイート数は相関係数は 0 に近く互いに独立であるが，図から，フォロワー数が 0～1000 人の間ではフォロワー数が多くなるにつれてツイート数の移動平均値が増加しているといえる．したがってツイート数とフォロワー数はある領域に限り，相関関係が全くないとはいいきれない．

次にデータセット A におけるフォロー数，ツイート数の散布図を図 4.15 に，データセット B におけるフォロー数，ツイート数の散布図を図 4.16 に示し，フォロー数に対するツイート数の関係性を分析する．

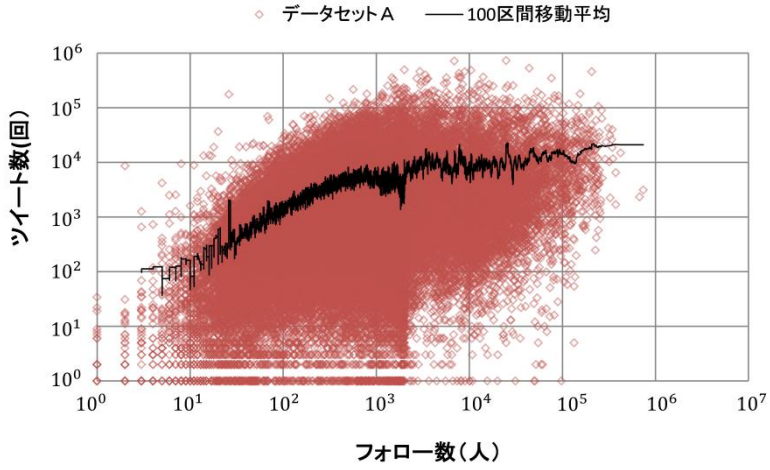


図 4.15：フォロー数に対するツイート数（データセット A）

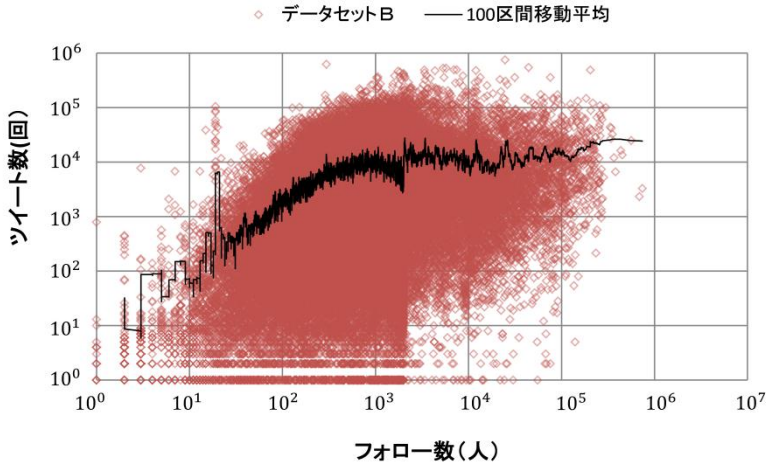


図 4.16：フォロー数に対するツイート数（データセット B）

フォロワー数とツイート数同様に，フォロー数とツイート数の相関係数も 0 に近く，互いに独立である．しかしフォロー数が 0~1000 人の範囲内ではフォロー数が多くなるにつれてツイート数の移動平均値が増加している．したがって，フォロー数とツイート数の間にはある領域に限り相関関係が全くないとはいいきれない．

次にデータセットAにおけるフォロワー数，フォロー数の散布図を図 4.17 に，データセットBにおけるフォロワー数，フォロー数の散布図を図 4.18 に示し，フォロー数に対するフォロワー数の関係性を分析する．

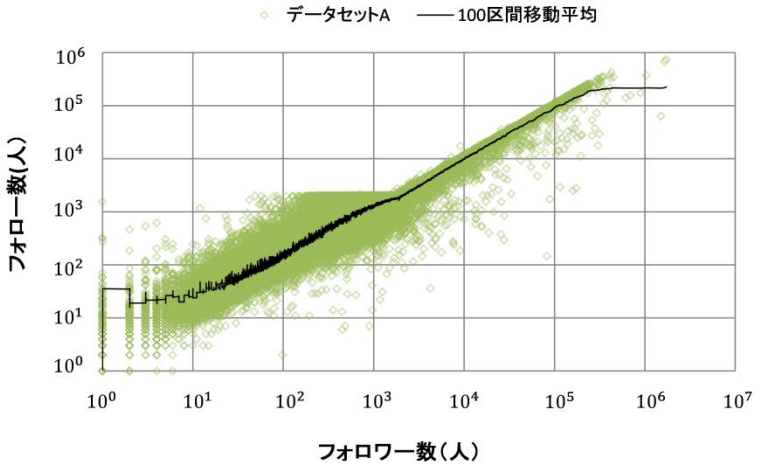


図 4.17：フォロワー数に対するフォロー数（データセットA）

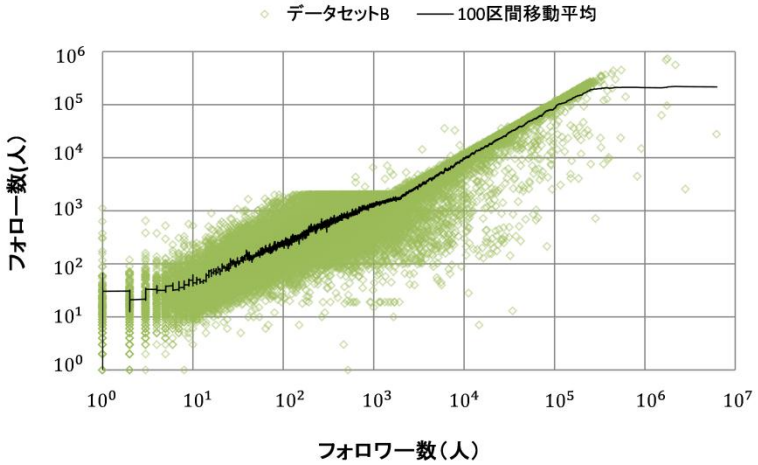


図 4.18：フォロワー数に対するフォロー数（データセットB）

データセットAのフォロワー数，フォロー数の相関係数は 0.8739，データセットBの 0.5259 とフォロワー数とフォロー数の間には正の相関がある．図を見てもその分布は正の傾きを持った直線関係を示しており，フォロー数とフォロワー数の間には強い正の相関があるといえる．また，フォロー数が 2000 人の部分で一定の値になっているが，これは 2000 人のフォロワーがいないとそれ 2000 人以上フォローできないというルールが影響している．

また，フォロワー数に対するフォロー数の関係を，ツイートの回数ごとに色分けして図 4.19（データセット A）図 4.20（データセット B）に示した．

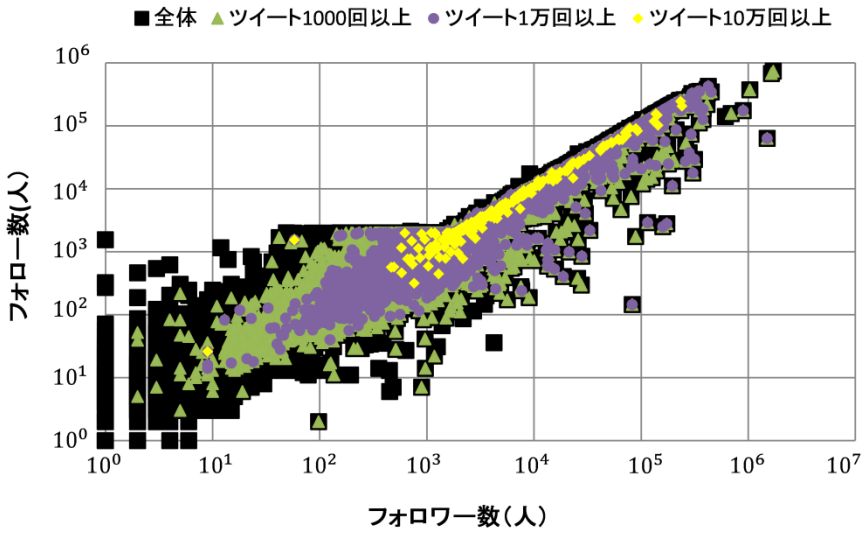


図 4.19：ツイート数に対するフォロワーとフォロー数の関係（データセット A）

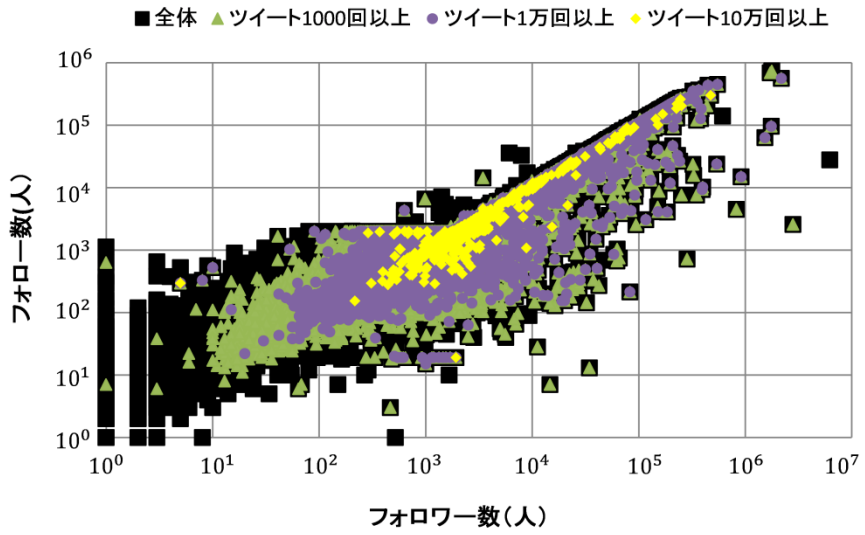


図 4.20：ツイート数に対するフォロワーとフォロー数の関係（データセット B）

ツイート数とフォロー数，ツイート数とフォロワー数の間には相関関係が見られなかった．互いに正の相関を持つフォロー数とフォロワー数の分布に，ツイート数が 1000 回以上の人，1 万回以上の人，10 万回以上の人を色分けした図を見ると，ツイート数が大きくなるにつれてフォロー数とフォロワー数はより強い直線関係をもつことがわかる．このことから，ツイート数はフォロー数とフォロワー数の相関関係を強める効果がある可能性を示唆している．

4.4　クラスター係数

4.4.1　定義

クラスター係数とは，近隣ノード間の接続関係の緊密さを示す指標である．一般的な定義としてノード a がノード b，ノード c とエッジで繋がっている場合に，ノード b とノード c がエッジで繋がっているかどうかという考え方で，クラスター係数が高い友人関係のネットワークでは，友達の友達は友達である可能性が高いといえる．

ノード v_i のクラスター係数を C_i とする． v_i の次数を k_i とすると， k_i 個ある i の隣接点から任意の 2 つの点を選び出すペアの個数は ${}_{k_i}C_2 = \frac{k_i(k_i-1)}{2}$ 個となる．それぞれのペアがエッジで繋がっている数を E_i とすると， C_i は以下の式で定義される．

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{3}$$

クラスター係数 C は，n 個ある頂点の C_i を平均した値で以下の式で定義される．

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{4}$$

では，図 4.21 のネットワークからクラスター係数 C を算出してみるとする．ノード a の隣接ノードは b，c，d，である．したがって（b，c），（c，d），（d，b）の 3 つのペアができる．3 つのペアの内，エッジで繋がっているのは（b，d）の 1 ペアである．したがって C_a=1/3 となる． C_b，C_c，C_d，C_e についても同様に算出し，それぞれ 1，0，1/3，0 となる．したがって C =5/18 と計算される．

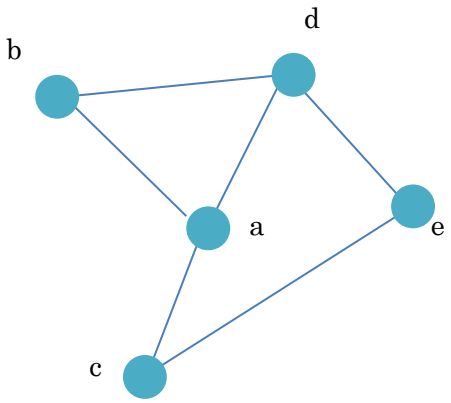


図 4.21：ノード a,b,c,d,e からなるネットワーク

4.4.2　twitter のネットワークにおけるクラスター係数

Twitter は 2 つのノード間にフォローとフォロワー2 方向のリンクが存在する．本稿では A が B をフォローしている場合 A がアウトリンクしており，B がインリンクを持っていると記述する．また，アウトリンクの数は出次数，インリンクの数は入次数とする．ここでいう出次数はフォロー数，入次数はフォロワー数に値するが，分析結果を示すにあたって，ネットワーク内のリンクと外側に伸びているリンクを区別するために，ネットワーク外に伸びているアウトリンク，インリンクを含めたリンク数はそれぞれフォロー数，フォロワー数とする．そこで，本稿では twitter のクラスター分析をクラスター係数もアウトリンクとインリンク両方に対してそれぞれ算出する．

前項の式(3)式(4)から，有向グラフである twitter のクラスター係数の算出方法を定義する．ある有向グラフ内のノード A，ノード B が互いにアウトリンクで繋がっている場合，アウトリンクで繋がっている数は（A→B），（B→A）の 2 つとなる．したがって あるノード v_i の出次数を k_i とすると，k_i 個ある，i の隣接点から任意の 2 つの点を選び出す場合も 2 通りを考慮し，その数は ${}_{k_i}C_2 = k_i(k_i - 1)$ 個となる．これにより，ノード v_i のアウトリンクに対するクラスター係数を $C_{i(out)}$ とする以下の式で定義される．

$$C_{i(out)} = \frac{E_i}{k_i(k_i - 1)} \tag{5}$$

アウトリンクのクラスター係数 C_{out} は，n 個ある頂点の $C_{i(out)}$ を平均した以下の式で定義される．

$$C_{out} = \frac{1}{n}\sum_{i=1}^{n} C_{i(out)} \tag{6}$$

では，あるノード a，b，c，d，e がアウトリンク，インリンクを持つ場合のネットワーク図を示した図 4.22 よりノード a のアウトリンクに対するクラスター係数 $C_{a(out)}$ およびノード a のインリンクに対するクラスター係数 $C_{a(in)}$ を算出してみるとする．

ノード a がアウトリンクで繋がっている隣接ノードは b，c，d であり，それぞれがアウトリンクで繋りうる組合せは（b→d），（b→c），（c→b），（c→d），（d→b），（d→c）の 6 つである．その内実際にアウトリンクで繋がっているのは（b→d），（d→b）の 2 つである．したがって $C_{a(out)} = \frac{1}{3}$ である．また，ノード a がインリンクを持つ（a をアウトリンクしている）隣接ノードは c，e であり，それぞれがインリンクで繋を持つ組合せは（c←e）（e←c）の 2 つである．その内実際にインリンクを持つのは(e→c)の 1 つである．したがって $C_{a(in)} = \frac{1}{2}$ である．

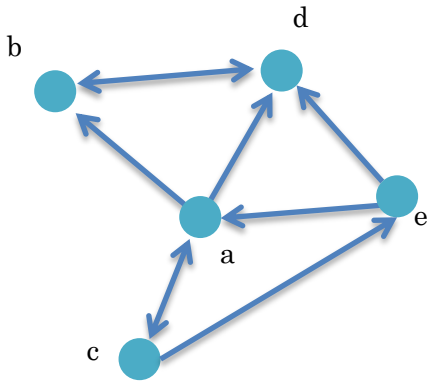


図 4.22：ノード a,b,c,d,e からなるネットワーク

以上の定義に従い，データセット C のクラスター分析を行う．データセット C は第 3 章で述べたように，自身のツイッターアカウントである「@sarashi_ya」を始点にフォロワーをたどり取得した 7713 ユーザーそれぞれのフォロワー一覧 ID 計 4713 万 9638 個取得したものである．各フォロワー一覧 ID の中に 7713 人の各ユーザーが含まれているかどうかを調べることで，7713 人のユーザーのフォロー，フォロワーの関係（アウトリンク，インリンクの繋がり）を抽出し，表 4.4.1 のようにまとめた．

	ユーザーa	ユーザーb	ユーザーc	ユーザーd	ユーザーe
ユーザーa	0	1	0	1	1
ユーザーb	0	0	0	1	0
ユーザーc	1	0	0	0	1
ユーザーd	0	1	1	0	0
ユーザーe	1	0	0	1	0

表 4.6：ユーザーごとのフォロー，フォロワーの繋がり

表 4.6 は，縦軸のユーザーのフォロワー一一覧 ID に横軸の ID が含まれていれば 1 を，含まれていなければ 0 の値を横軸にとっていったものである．したがって横軸を見るとそれぞれのアウトリンクしている人が 1 で示され，縦軸を見るとそれぞれのインリンクをもつ（アウトリンクされている人）が 1 で示される．また，横軸の 1 の合計数が出次数，縦軸の 1 の合計数が入次数となる．

表 4.7 に算出したアウトリンクのクラスター係数C_{out}，インリンクの平均クラスター係数C_{in}，それぞれの値と最小値，最大値，を記す.また表 4.8 に出次数，入次数の最小値，最大値，平均値，合計値を記す.

表 4.7： クラスター係数C_{out}，クラスター係数C_{in}の最小値，最大値，平均値

	最小値	**最大値**	**平均値**
C_{out}	0	1	0.3542
C_{in}	0	1	0.3165

表 4.8：出次数，入次数の最小値，最大値，平均値，合計値

	最小次数	最大次数	平均次数	次数合計
出次数	1	2076	124	958413
入次数	0	2100	124	958413

アウトリンクの平均クラスター係数は 0.3542 で，インリンクの平均クラスター係数は 0.3165 であった．これは，両者ともに mixi のクラスター係数 0.237[8]よりも高い結果となった．また，アウトリンクとインリンクの平均クラスター係数はやや誤差はあるものの，ほぼ同じであるということが分かる．

出次数の最小値が 1 であるのは，データの取得方法に起因している．各ノードは，フォロワーをたどってピックアップしたものであり，アウトリンクを最低 1 つは保有しているノードとなる．出次数，入次数の最大次数を持つユーザーは同一人物であり，「@euro_tour ヨーロッパ旅行情報部」というヨーロッパ情報を流すアカウントであった．

次に，個々のクラスター係数を条件ごとに比較し，クラスター係数と次数，フォロー数，フォロワー数，ツイート数との関係性を分析していく．

まず，図 4.23 よりノードごとのアウトリンクのクラスター係数に対するインリンクのクラスター係数の値を比較しその関係性を分析する．

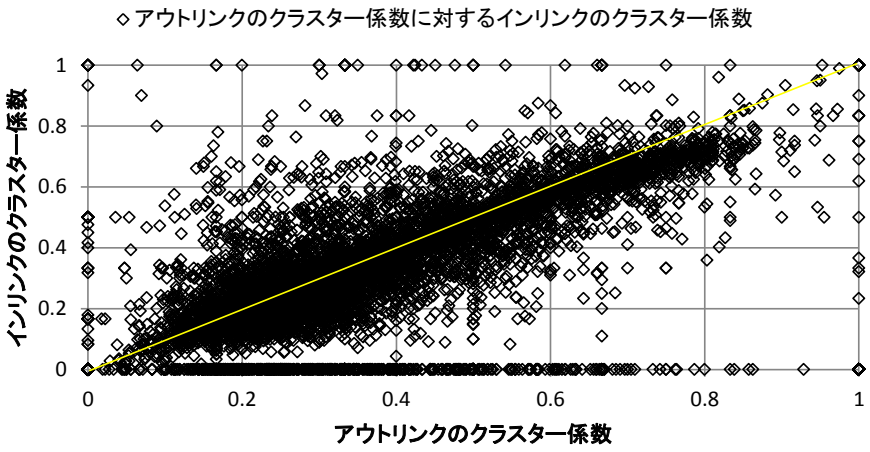


図 4.23：アウトリンクとインリンクに対するクラスター係数値の比較

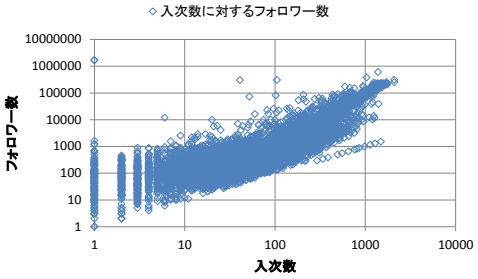


図 4.24：フォロワー数に対する入次数

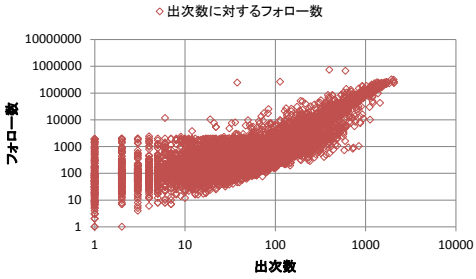


図 4.25：フォロー数に対する出次数

アウトリンクのクラスター係数, インリンクのクラスター係数の 2 つ値の間には直線関係があり，相互に関係し合っているといえる．また，図 4.24 と図 4.25 からフォロー数と入次数，フォロワー数と出次数の大きさは範囲が大きいが直線的に増加している．したがってフォロー数の大小は入次数の大小に, フォロワー数の大小は出次数の大小に少なからず関係しているといえるであろう．

次に，図 4.26 から出次数に対するクラスター係数の値，図 4.27 から入次数に対するクラスター係数の値を比較する．

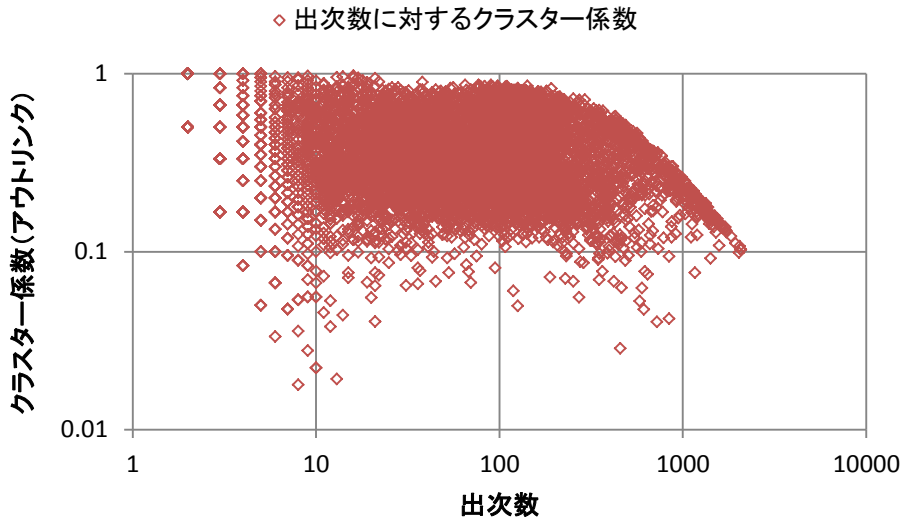


図 4.26：出次数に対するアウトリンクのクラスター係数

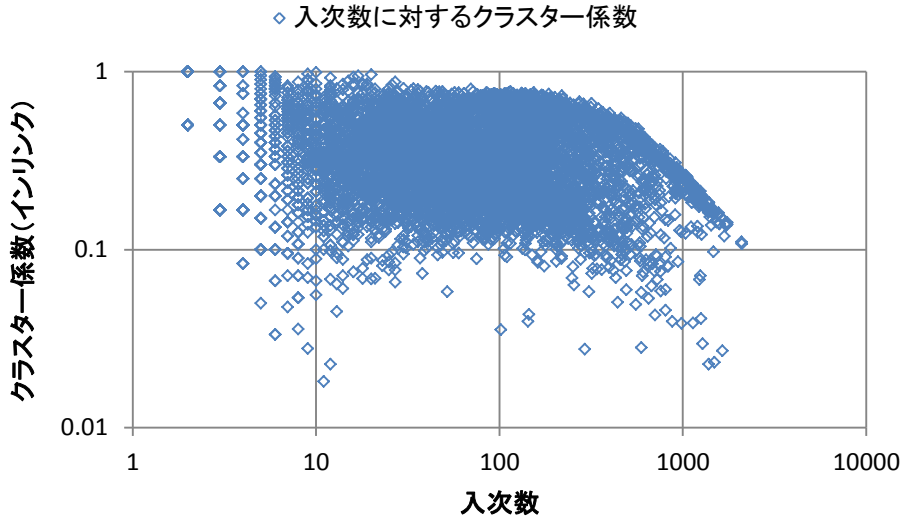


図 4.27：入次数に対するインリンクのクラスター係数

出次数は 100 を境目に，入次数は 300 を境目に，次数が大きくなるにつれてそのクラスター係数の値は徐々に小さくなる傾向が目にとれる．これは，次数が大きければ大き

いほど，幅広いジャンルの人とフォロー，フォロワーの関係で繋がっているからであると考えられる．

次に図 4.28，図 4.29 からフォロー数（出次数）に対するクラスター係数の値，フォロワー数（入次数）に対するクラスター係数の関係を読み取る．

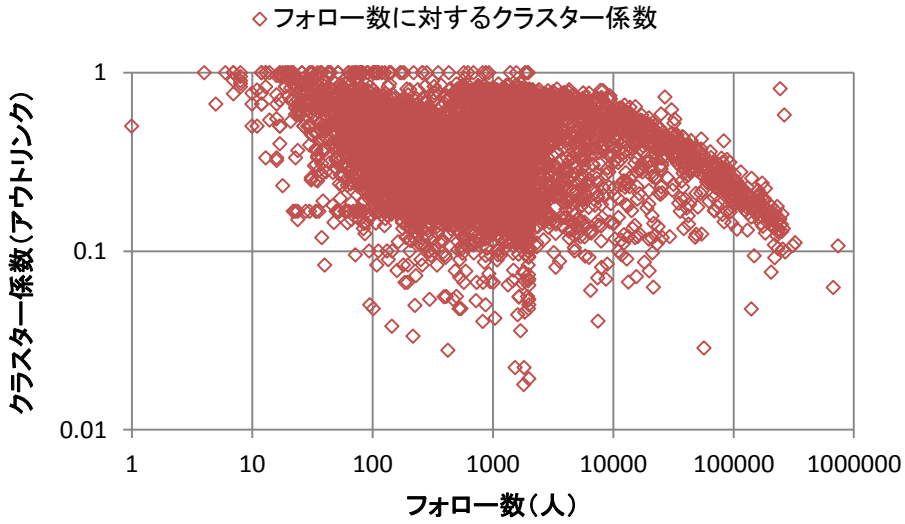


図 4.28：フォロー数（出次数）に対するクラスター係数（アウトリンク）の関係

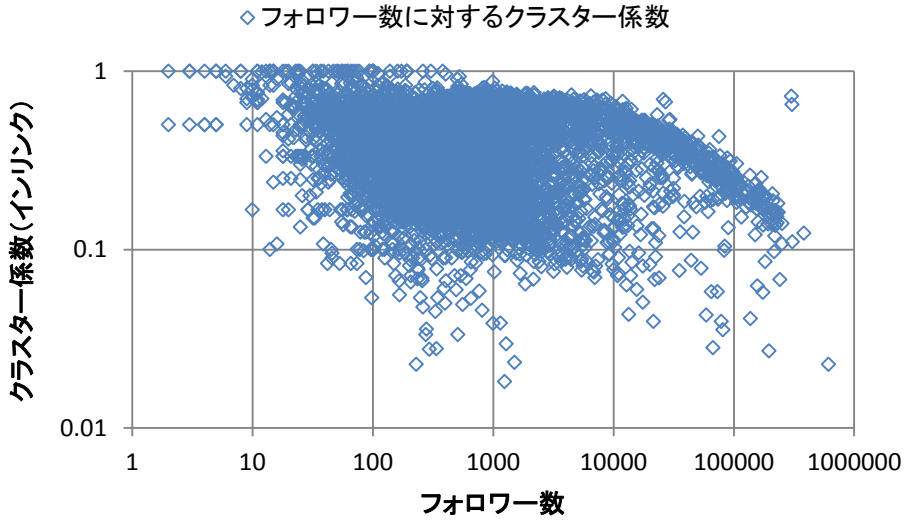


図 4.29：フォロワー数（入次数）に対するクラスター係数（インリンク）の関係

フォロワー数（入次数），フォロー数（出次数）ともに，5000 人まではそのクラスター係数も幅広く分布している．しかし 10000 人を境目にフォロー数，フォロワー数が

多くなるにつれて，徐々にクラスター係数の値は小さくなっていく傾向にある．この傾向は出次数，入次数とクラスター係数の関係と類似しており，フォロー数と出次数，フォロワー数と入次数が互いに関係し合っているという可能性を示唆している．

次に図 4.30，図 4.31 からツイート数に対するアウトリンクのクラスター係数の値，ツイート数に対するインリンクのクラスター係数の関係を読み取る．

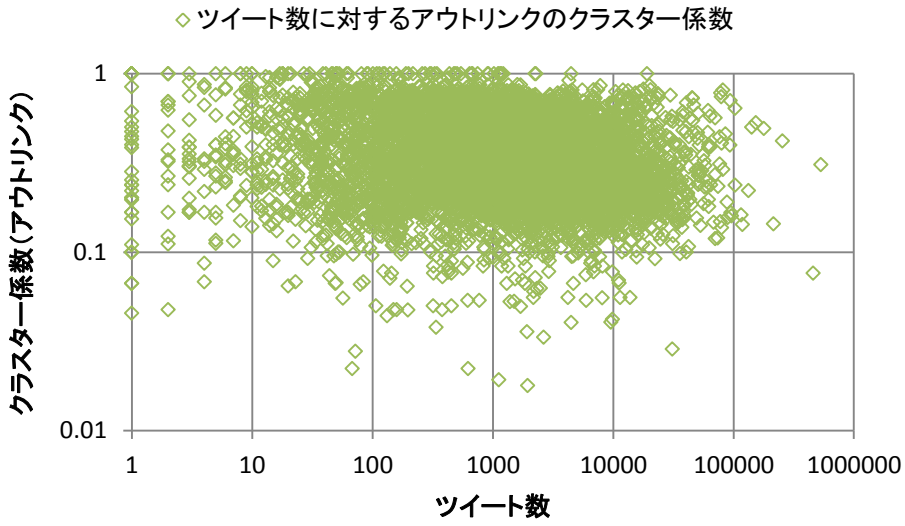


図 4.30：ツイート数に対するクラスター係数（アウトリンク）

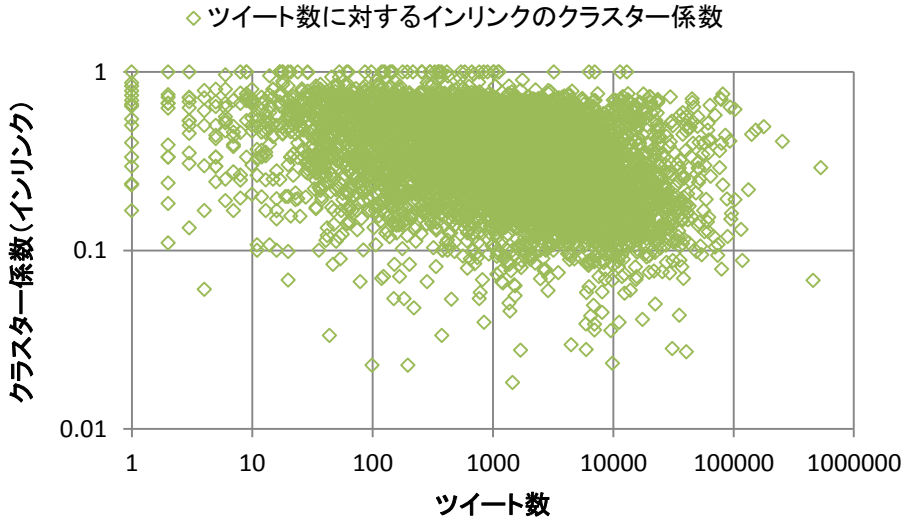


図 4.31：ツイート数に対するクラスター係数（インリンク）

ツイート数の大小に関わらずアウトリンク，インリンク両者のクラスター係数は幅広く値をとっており，ツイート数の大小はクラスター係数の大小に影響を及ぼさないといえるであろう．

第 5 章　まとめ

本研究では twitter のソーシャルグラフを中心に分析を行った．その結果次数分布からは, twitter のネットワークはスケールフリー性があるとは断定しがたいということ, 相関係数, 散布図からはフォロー数が多いほどフォロワー数は多くなりツイート数の大小はフォロー数フォロワー数に影響を及ぼしているとはいえないが, フォロー数とフォロワー数の相関関係に対して影響を及ぼしているということ，クラスター分析からは twitter が高い平均クラスター係数を持つ一方で次数が高くなればなるほどその値は低くなっていくという関係性があるということが分かった．しかし，分析で使用したデータは最大 10 万人であり，そのデータの取得方法も 2 種類のみであったため，現在における twitter の利用者数を考慮するとデータ数が十分であると断定できない．

今後は，引き続きユーザー情報の取得を行い分析の精度を高めるとともに，ネットワーク上の平均経路数や同類選択性といった指標からさらなるソーシャルグラフの分析を進めていく．また，データの取得方法をパターン化し，各個人のツイート内容の取得や，ツイート，リツイートによる情報伝搬の追跡から twitter の情報伝搬経路を分析していく．

参考文献

[1]B. Krishnamuthy，P. Gill, and M. Arilitt. ”A Few Chirps About Twitter” ACM WOSN’08, 2008.

[2]A.Mislove, Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. “Measurement and analysis of online social networks”. ACM IMC, 2007.

[3]A. Java, X. Song, T. Finin, and B. Tseng. “Why we twitter: Understanding microblogging usage and communities”. KDD, 2007.

[4]H. Kwak, C. Lee, H. Park, and S. Moon. “What is twitter, a social network or a media?”. ACM WWW,2010.

[4]Twitter:What are you doing? http://www.twitter.com

[5]山本裕介，“Twitter API ポケットリファレンス，”技術評論社，2011.

[6] Oauth Community Site：http://oauth.net/.

[7] http://dev.twitter.com/pages/auth.

[8]丸井淳己，加藤幹生，松尾豊，安田雪，“mixi のネットワーク分析” 情報処理学会第 72 回全国大会講演論文集，No2, pp.553-554，2010

謝辞

本研究におきまして，知識経験に乏しく至らない点の多い私を，終始暖かくご丁寧に指導していただきました塩田茂雄教授に深く感謝しお礼申し上げます．また，日々共に支え合い励まし合った塩田研究室の同輩，多くの助言をしていただきました塩田研究室の先輩の方々に深く感謝致します．ここに感謝の意を表します．誠にありがとうございました．