3A2-3

mixi のネットワーク分析

Network Analysis of mixi

丸井 淳己[*1] 加藤 幹生[*2] 松尾 豊[*3] 安田 雪[*4]
Junki Marui Mikio Kato Yutaka Matsuo Yuki Yasuda

[*1]東京大学 The University of Tokyo
[*2]株式会社ミクシィ mixi, Inc.
[*3]東京大学 The University of Tokyo
[*4]関西大学 Kansai University

Online advertising has assumed a new aspect with the appearance of Google AdSense of which content is relevant to the website. Our purpose is to propose a strategy of online advertising in a new generation, more specifically, advertising relevant to the relations of people. To achieve this, first, we investigated the structure of the personal networks developed in mixi, a Japanese social networking service (SNS), and discovered the ways how information travels on homophily network. Second, we classified relations of people into several types by defining the attributes of personal links and using k-means clustering. Finally, we investigated the network motif and the co-occurrence of the link type.

1. はじめに

ソーシャルネットワーキングサービス (SNS) は今や多くの人にとって欠かせないものとなっている． SNS を特徴付ける性質の一つにスモールワールド性があり，情報推薦や口コミによる広告媒体としての大きな可能性を秘めている．

我々は今回日本で最大の SNS である mixi の分析をした．本研究で使用したデータは 2009 年 5 月時のものであり，ユーザ数は 16,937,041，リンク数は 414,250,844 本であった．リンク先の友人を mixi では「マイミクシィ」(以下略してマイミク) と呼称している．ユーザのデータは性別・年齢・アクティブ度 (最終ログイン日からの日数)・都道府県で構成され，ID はシャッフルされているためユーザの特定は出来ないようになっている．さらに「あしあと」と呼ばれるユーザ間のアクセス履歴 (2009/5/1-5/30, 秒単位での時刻付) 1,522,157,737 レコードの提供を受けた．

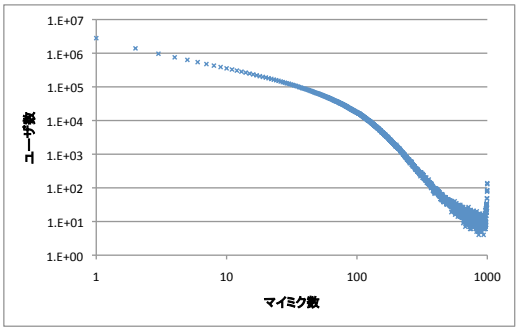


図 1: 次数分布

2. mixi ネットワークの性質

平均マイミク数は 24.46 人で，次数分布は図 1 のようになっている．次数 1000 の付近でユーザ数が増えているのは，マイミク数が現時点で 1000 に制限されているためだと考えられる．クラスタ係数は 0.237 であり 2006 年に行われたネットワーク分析での結果 (0.328) と比較するとこの値は小さい [松尾 07].

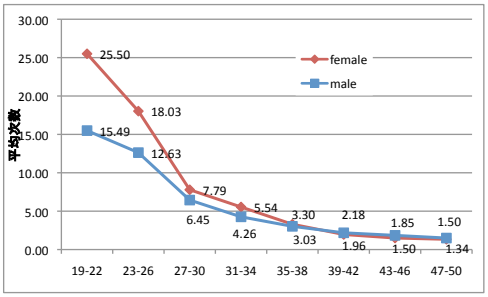


図 2: 平均次数

平均経路長は全体の 2.04%ほどの 340,717 人に計算をしたところ 5.457 であった．この値は 2006 年での結果 (5.53) と比べて大差ない結果である．

最初の分析として，ユーザの属性 (年齢・性別) で切り分けてネットワーク分析を行った．同じ性別と年代である人々だけを構成員とした同質集団のネットワークは表 1 のようになっている．各ネットワークの平均次数は図 2 に示した．

表 1: 各集団の基礎データ

年代	女性		男性	
	ノード数	リンク数	ノード数	リンク数
19-22	1,947,322	49,660,788	1,600,907	24,804,214
23-26	1,933,718	34,867,346	1,635,141	20,643,780
27-30	1,485,059	11,567,942	1,308,112	8,432,692
31-34	1,132,498	6,279,684	993,353	4,233,516
35-38	744,817	2,455,068	711,797	2,153,468
39-42	398,438	782,186	439,918	959,652
43-46	200,967	302,028	248,506	458,788
47-50	118,237	158,134	149,587	224,024

男女ともに年代を上がるほど平均次数が下がっていく．さらに図 3 と図 4 の比較でも分かるように、若い年代は同世代へ多

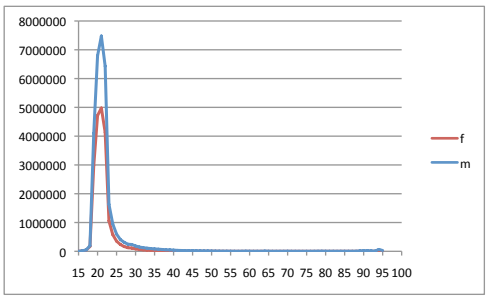


図 3: 19-22 歳男性のリンク先の年齢・性別分布

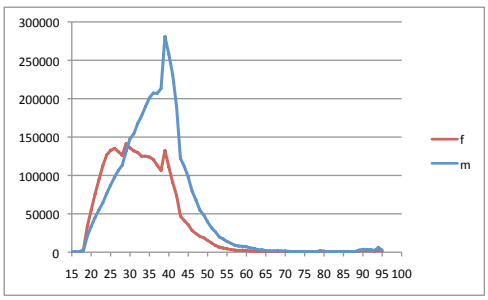


図 4: 39-42 歳男性のリンク先の年齢・性別分布

くリンクを張るのに対して年代を上がると違う世代へのリンクの割合が増す。例えば 39 歳から 42 歳の男性の場合，女性へのリンクは同年代よりも 20 代の方が多いことがわかる．リンクの双方の次数の相関係数である Assortativity Coefficient(以下 AC) についても計算を行った (図 5)．これを見ると，女性が平均初婚年齢 (28.3 歳) を境に大きく変化している事がわかる．

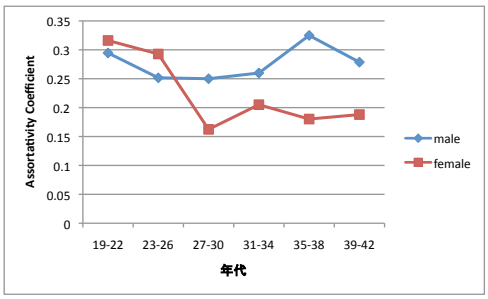


図 5: Assortativity Coefficient

以上の分析により年代が上がるほど相対的に他の年代・性別といった属性の違う人へのつながりが多くなり，それぞれ人間関係が違う性質を持つことが推測出来る．そこで次の分析では同質に絞らずに全ての人間関係を種類分けした．

3. マイミクの分類

クラスタリングによるリンクの種類分けを行った．リンクの属性を両端のユーザの年齢差，性差 (同性:0, 異性:1)，次数差，共通隣人数 (全て絶対値) の 4 つで定義した．各属性をそれぞれの平均を引いて標準偏差で割る正規化後，k-means 法によるクラスタリングを行った．適切な k(クラスタ数) を設定するために以下に示した Clustering Quality という指標を用いた．これは Bollegala らが階層化クラスタリングの際に用いたものである [Bollegala 06]．まず，あるクラスタ Γ 内での凝集度合いは式 1 で与えられる．

$$C(\Gamma) = \begin{cases} 1 & |\Gamma| = 1 \\ \dfrac{2}{|\Gamma|(|\Gamma|-1)} \displaystyle\sum_{u\in\Gamma} \sum_{v\in\Gamma, v\neq u} sim(u,v) & \text{otherwise.} \end{cases} \quad (1)$$

この式の u, v は同クラスタにあるマイミクリンクの属性ベクトルである．sim として今回はコサイン類似度を用いている．

この式からクラスタ内の凝集度合いのスコアは式 2 で表される．Λ はクラスタ Γ の集合で，今回は k-means 法で k 個に分けたクラスタの集合を表している．この式の意味する所は各クラスタ内の凝集度合いの平均である．

$$\mathrm{IntCor}(\Lambda) = \frac{1}{k} \sum_{\Gamma\in\Lambda} C(\Gamma) \quad (2)$$

さらにクラスタ間の離れ具合のスコアは式 3 で表される．これはある 2 つのクラスタを結合させたときに凝集度合いが一番小さくなるような 2 つのクラスタペアを選び，その二つのクラスタ類似度を 1 から引いた値としてクラスタ間の離れ具合をスコアリングしている．

$$\mathrm{ExtCor}(\Lambda) = 1 - \frac{1}{|\Gamma_a||\Gamma_b|} \sum_{u\in\Gamma_a} \sum_{v\in\Gamma_b} sim(u,v) \quad (3)$$

$$(\Gamma_a, \Gamma_b) = \arg_{\Gamma_i,\Gamma_j\in\Lambda} \max C(\Gamma_i \oplus \Gamma_j) \quad (4)$$

(式 4 の $\oplus$ は，クラスタを結合させると言う意味である．)
以上より Clustering Quality は式 5 として定義出来る．

$$Q(\Lambda) = \frac{1}{2}(\mathrm{IntCor}(\Lambda) + \mathrm{ExtCor}(\Lambda)) \quad (5)$$

この値を $k = 3$〜10 で計算すると図 6 のようになって $k = 5$ が最大である．このときクラスタリングによって表 2 のような中心ベクトルが求まる．各クラスタの中心ベクトルの特徴から年齢差，次数差，仲良し，同性同年代，異性リンクとラベルをつけた．表 2 にのせたアクセス数はそのリンクを通る 30 日間にわたるアクセスの平均である．

表 2: クラスタの中心ベクトル

Cluster	年齢差	性差	次数差	共通隣人	アクセス
1	25.14	0.32	62.30	4.29	2.24
2	5.76	0.45	403.95	4.92	1.43
3	1.58	0.16	65.58	24.45	2.62
4	1.38	0.00	39.44	3.87	1.96
5	2.46	1.00	51.19	4.75	2.04

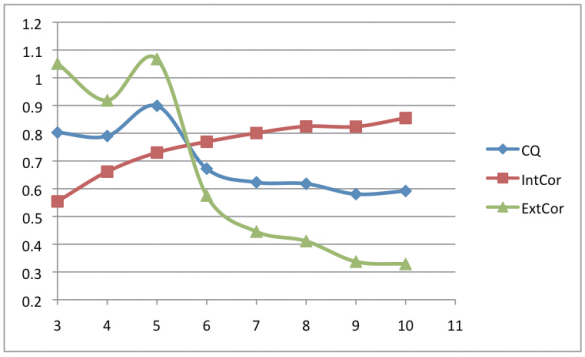


図 6: Clustering Quality

それぞれのリンククラスタは簡単に次のような特徴を持つ.

クラスタ 1 年齢差リンク

- 年齢の離れた者同士をつなぐリンク.

クラスタ 2 次数差リンク

- マイミク数の離れた者同士をつなぐリンク. 比較的高い年齢差も伴う.

クラスタ 3 仲良しリンク

- 共通隣人数が多い者同士をつなぐリンク. 年齢差が低くやや同性である傾向が見られる.

クラスタ 4 同年代同性リンク

- 同性かつ年齢差の低い者同士をつなぐリンク. ただし仲良しリンクと違って共通隣人数が低い.

クラスタ 5 異性リンク

- 異性同士をつなぐリンク. 年齢差を 2 程度伴う.

各リンクの種類によって情報の伝播の仕方が変わると推察される. 例えば年齢差のリンクを多く持つユーザに同年代しか興味を持たないような情報を流しても伝播が期待できない. さらにどの種類とどの種類のリンクは同じような情報を通すのかを考えるためにはこのリンクの種類の共起を考えると良い. 三者関係を取り出してどの種類のリンクでお互いが繋がっているか, 行き来しているかを調べた.

表 3: 色とリンク種別の対応

色	クラスタ
	1. 同年代同性リンク
	2. 年齢差リンク
	3. 仲良しリンク
	4. 異性リンク
	5. 次数差リンク

表 4: マイミクリンクのトライアド

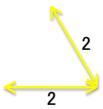
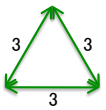
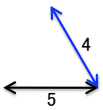
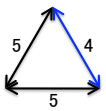
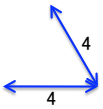
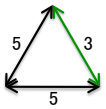
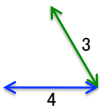
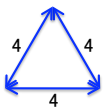
順番	リンク 2 本		リンク 3 本	
	トライアド	頻度	トライアド	頻度
1		15.91%		0.928%
2		15.47%		0.326%
3		11.48%		0.301%
4		9.33%		0.262%

表 5: あしあとのトライアド

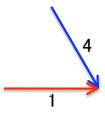
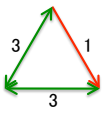
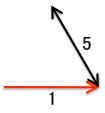
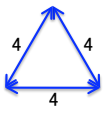
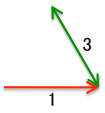
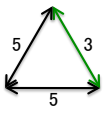
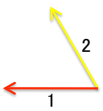
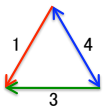
順番	リンク 2 本		リンク 3 本	
	トライアド	頻度	トライアド	頻度
1		14.71%		0.6640%
2		12.90%		0.0998%
3		12.37%		0.0861%
4		9.36%		0.0816%

表 4 はトライアドを三者間にリンクが 2 本の場合と 3 本の場合で分けて頻度順に並べたものである. エッジに書かれた数字及び色は表 3 と対応している. あしあとの場合は表 5 に示した.

この共起を見ると, マイミクでは次数差リンク二つというパターンが多く出現するのに対してあしあとでは同性同年代リンク 2 本をお互いに行き会うパターンが最も多かった. しかしながら必ずしも属性の似た人同士がつながっている訳ではなく, 異性リンク二つといったパターンが多く, ある性に特化した情報は流れにくい事が推察される. あしあとのトライアドはマイミクのそれとは異なると言える. さらに比較的マイミクの頻度の高いトライアドが同質なリンク (同性同年代リンク・仲

良しリンク) が多かったのに対して，あしあとでの頻度の高いトライアドは年齢差リンクと異性リンクといった同質性以外同士のリンク共起が見られる事も特徴的である．

リンク分類によって 5 種類のうち 2 種類は同質なものをつなぐリンクであり，アクセスの多さや本数の多さで情報伝播を考えたときにも主要な役割を果たすと考えられるが，2 点間だけでなく 3 点間のインタラクションを見ると同質な者を結ぶリンクだけが共起する訳ではなく，同質ではないリンクが情報伝播の際に大きく影響する事が分かった．さらにアクセスした関係のトライアドを見ると，非同質なグループへのリンクがさらに重要性を増すことが分かった．

4. まとめ

本研究では mixi のデータを対象として三つの分析を行なった．一つはネットワーク全体の議論で，マイミク・あしあと共に基礎的な指標により性質を明らかにした．2 つ目に同質集団のネットワーク分析を行い，若年層は同質な集団の中で非常に強く繋がり年齢が上がる毎に人間関係の幅が広がってゆく事を明らかにした．特に中年は若い層に多くリンクを張っているが，アクセスは同年代がマジョリティーを獲得している事が分かった．また AC の値から 27 歳以前と以降での女性の人間関係の構造的な違いが示唆された．次にマイミクリンクの種類分けをし，クラスタリングクォリティの最も高い 5 種類に分けると良い事を示し，さらにその 5 種類の中心ベクトルを見る事で各クラスタの特徴を明らかにした．最後にそれらマイミクリンクの共起性を調べるために Network motif 分析を行ない，マイミクとあしあとでトライアドの発生パターンが違う事，マイミクよりもあしあとは異質な存在同士を結んだトライアドが多く出現する事が分かり，同質への情報伝播だけを着目すると多くの機会を損失しうることを示した．

これら三つの分析はそれぞれ情報の伝播について知見を与えるものであり，SNS は一手に人間関係のデータを握っているのでこれをうまく活用する事が出来れば SNS としてのブレイクスルーを引き起こせるばかりか，情報推薦システムや広告にとっても大きな一歩となる．本研究がそのような SNS の眠っていた価値を活かす基礎に繋がれば著者らの幸いとする所である．

参考文献

[Bollegala 06] Bollegala, D., Matsuo, Y., and Ishizuka, M.: Disambiguating Personal Names on the Web using Automatically Extracted Key Phrases, in *Proceeding of the 2006 conference on ECAI 2006*, pp. 553–557, Amsterdam, The Netherlands, The Netherlands (2006), IOS Press

[松尾 07] 松尾 豊, 安田 雪：SNS における関係形成原理 -mixi のデータ分析-, 人工知能学会論文誌, Vol. 22, No. 5, pp. 531–541 (2007)