DEIM Forum 2013 D10-5

レビューサイトにおけるユーザ間類似度分析

山岸　祐己† 　斉藤　和巳†

† 静岡県立大学経営情報イノベーション研究科　〒 422–8526 静岡県静岡市駿河区谷田 52-1
E-mail: †{j12115,k-saito}@u-shizuoka-ken.ac.jp

あらまし　インターネット上のレビューサイトでの口コミ情報などは，オンライン環境での購買行動を始めとし，ユーザ間で多様な活動に影響を及ぼしている．具体的には，多くのユーザから信頼されているユーザにより，ある商品に対して高評価のレビューが投稿されれば，周囲のユーザ，特に類似度が高いユーザに対し購買意欲を引き起こさせ，結果その商品の販売が大幅に促進されるようなことも起こり得る．よって，レビューサイトにおけるユーザ間の類似度の定式化や，類似度の時系列的変化の観測は，ウェブ情報学において極めて重要といえる．本研究では，ユーザ間の類似度が時間とともに変化する動的な側面を考慮し，現実の大規模レビュー時系列データを用いて統計的分析を行う．
キーワード　レビューサイト，ユーザ間類似度，時系列分析，K-median

Similarity Analysis Between Users in Online Review Sites

Yuki YAMAGISHI† and Kazumi SAITO†

† Graduate School of Management and Information of Innovation, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka 422–8526, Japan
E-mail: †{j12115,k-saito}@u-shizuoka-ken.ac.jp

Abstract　The word of mouth information of online review sites on the internet are affecting various activities from person to person. For example, if a user who is trusted by many users posted a high rating review to a product, around the users, especially highly similar users will have interest in the product. As a result, the sales of the product may be increased significantly. Therefore, the formularization of the similarity between users in online review sites, and observation of change of the similarity are very important research. In this research, we analyse the large-scale review time series data, in consideration of the dynamic side in which the similarity between users changes with time.
Key words　online review sites, similarity between users, time series analysis, K-median

1. はじめに

レビューサイトとは，商品やサービスについてのレビューを投稿することができるウェブサイトの総称である．レビューは点数・文章・画像から成ることが多く，レビュー点数の平均点が，アイテムに対する一般的な評価指標として扱われている．レビューサイトについては，既に多様な分析や研究が展開されている [1]．近年，レビューサイトにおけるユーザーのレビュー行動が非常に活発であり，それらロコミ情報がオンライン環境での購買行動を始めとし，ユーザ間で多様な活動に影響を及ぼしているため，サイトそのものが商品やサービスのプロモーションを左右する重要なメディアになりつつある．特に，投稿数が群を抜いている重要ユーザの高評価レビューは，対象商品の販売を大幅に促進させる可能性が高い．よって，消費者に宣伝と気付かれないように宣伝行為をするステルスマーケティングに関する注目も集まっている．

このようなインターネット上での商品に関する情報拡散や，どのような商品が好まれるかの意見形成の仕組みは，ウェブ情報学において極めて重要な研究対象であり，これらに対して一般的な数理モデリングの視点での研究が展開されている [2]．ただし，これらの研究では，ユーザ間の信頼や評判レベルは，どのペア間でも互いに一様と単純化してモデリングしているため，精緻な情報拡散や意見形成の分析には限界があると考えられる．よって，我々はユーザ間の信頼関係が時間とともに変化する動的な側面も考慮し，ソーシャルネットワーク上でのユーザ間の動的信頼メカニズムを数理モデリングする着想に至った．今回は，その動的信頼モデルを実現するための事前研究として，現実の大規模レビュー時系列データを用いた統計的分析を扱う．

本稿の構成は以下となる．まず，レビュー時系列データにおける動的レビュー類似度を定式化し，それらを K-median 法

によってクラスタリングする方法，及びクラスタリング結果の意味を検証する方法について説明する．そして，実験で用いたデータセットの詳細を説明し，実験結果とまとめについて述べる．

2. 分 析 手 法

分析手順と，その要素技術について説明する．ここで，対象データのユーザ集合を V，アイテム集合を $I(|I| = N)$，レビュー点数が取りうる整数値を $M = \{1, \cdots, m\}$，最も多くのレビューを投稿したトップユーザを $v_0 \in V$ とし，以下の手順により分析を行う．

（1） v_0 と $v \in V \setminus \{v_0\}$ の動的類似度ベクトル $\boldsymbol{y_v}$ を計算；

（2） $\boldsymbol{y_v}$ を K-median 法によりクラスタリング；

（3） クラスタ間のユーザ属性類似度を調べ，クラスタリング結果の意味を推定；

2.1 動的レビュー類似度

データ上の最古のレビュー投稿時刻を 0，最新のレビュー投稿時刻を T とすれば，時刻区間 $[s, s+\Delta](0 \leqq s \leqq T-\Delta)$ を定めることできる．この時刻区間において，ユーザ $v \in V \setminus \{v_0\}$ が，n 番目のアイテム $i_n \in I$ に対して付与したレビュー得点を $E(v, i_n) \in M$ とし，この区間の v のレビュー得点を要素とする N 次元ベクトルで

$$\boldsymbol{x}_v(s) = \{E(v, i_1), E(v, i_2), \cdots, E(v, i_N)\}^T, \tag{1}$$

と書き表せば，トップユーザ v_0 と v の類似度は

$$y_v(s) = \frac{\boldsymbol{x}_{v_0}(s)^T \boldsymbol{x}_v(s)}{||\boldsymbol{x}_{v_0}(s)|| \cdot ||\boldsymbol{x}_v(s)||}, \tag{2}$$

のように算出することができる．さらに，設定した全ての s における $y_v(s)$ を時刻順で要素に持つ動的類似度ベクトルを $\boldsymbol{y_v}$ とすれば，ユーザ u，v 間の動的類似度の類似度は

$$\rho(u, v) = \frac{\boldsymbol{y}_u^T \boldsymbol{y}_v}{||\boldsymbol{y}_u|| \cdot ||\boldsymbol{y}_v||}, \tag{3}$$

と求めることができる．

2.2 K-median クラスタリング

K-median（K-medoid とも呼ばれる）法は，非階層クラスタリングで有名な K-means 法と同様に，N 個のオブジェクト集合 $\mathcal{V}$ が与えられたとき，オブジェクト集合を K 個のクラスタに分割する手法である．任意のオブジェクトペア $u, v \in \mathcal{V}$ 間に，2.1 節でも用いた類似度 $\rho(u, v)$ を定義し，オブジェクト集合の中から他のオブジェクトとの類似度の和が高い代表オブジェクトを選定し，類似度の高いオブジェクトペアは同じクラスタに，類似度の低いオブジェクトペアは異なるクラスタに属するように分割する．一般的に，平均（mean）より中央値（median）の方が頑健であることが知られている．K-median の解法には反復法や貪欲法があるが，本研究では解の一意性が保証される貪欲法を採用する．さらに，貪欲解法の目的関数のサブモジュラ性より，厳密解ではないものの，ある程度妥当な精度で最悪ケースの解品質が理論的に保証されている [3]．貪欲法とは，既に選定した代表オブジェクトを固定し，ある評価関数値を最大にするオブジェクトを求め，目的関数が増加するならば代表オブジェクト集合に追加することで，結果の代表オブジェクト集合を求める方法である．各オブジェクトは，最も類似度の高い代表オブジェクトと同じクラスタに割り当てられる．既に選定した代表オブジェクト集合を $\mathcal{P}$ とし，新たに追加を試みるオブジェクトを w とするとき，本稿では，以下の目的関数を考える．

$$f(\mathcal{P} \cup \{w\}) = \sum_{v \in \mathcal{V}} \max\{\mu(v\,; \mathcal{P}), \rho(v, w)\}. \tag{4}$$

ここで，$\mu(v\,; \mathcal{P})$ は既に選定された代表オブジェクトとの類似度の最大値を表し，$\mu(v\,; \mathcal{P}) = \max_{w \in \mathcal{P}}\{\rho(v, w)\}$ で定義される．以下に貪欲法による K-median 法のアルゴリズムを説明する．

（1） $k \leftarrow 1$, $\mathcal{P}_0 \leftarrow \emptyset$，各オブジェクト $v \in \mathcal{V}$ に対し，$\mu(v\,; \emptyset) \leftarrow 0$ と初期化する；

（2） 式 4 で $\hat{p}_k = \arg\max_{w \in V \setminus \mathcal{P}_{k-1}}\{f(\mathcal{P}_{k-1} \cup \{w\})\}$ を求め，$\mathcal{P}_k \leftarrow \mathcal{P}_{k-1} \cup \{\hat{p}_k\}$ とする；

（3） $k = K$ ならば $\hat{\mathcal{P}}_K = \{\hat{p}_1, \ldots, \hat{p}_K\}$ を出力し終了する；

（4） 各オブジェクト $v \in \mathcal{V}$ に対し，$\mu(v\,; \mathcal{P}_k)$ を求め，$k \leftarrow k+1$ としステップ 2. へ戻る．

各オブジェクトを，最も類似度の高い代表オブジェクト $p_k \in \mathcal{P}$ のクラスタ $\mathcal{C}_k$ に割り当てる．

2.3 ユーザ属性類似度

ユーザ u，v 間の属性類似度を，数値の場合と名義値の場合とで別々に定義する．年齢のような数値属性の場合は，差の絶対値を類似度として用い，それを $AV_{u,v}$ とする．性別のような名義値属性の場合は，名義値属性からなる Z 次元ベクトルを $\boldsymbol{a}_v = (a_{v,1}, \cdots, a_{v,Z})^T$ として，以下で定義される $NV_{u,v}$ とする．

$$NV_{u,v}(\boldsymbol{q}) = \sum_{z=1}^{Z} q_z \delta(a_{u,z}, a_{v,z}). \tag{5}$$

ここで，$\boldsymbol{q}$ は，各属性ごとの類似度への寄与度を制御する M-次元のパラメータベクトルであり，$\delta(\cdot, \cdot)$ は，

$$\delta(x_{u,m}, x_{v,m}) = \begin{cases} 1 & \text{if } x_{u,m} = x_{v,m}, \\ 0 & \text{otherwise;} \end{cases} \tag{6}$$

により定義される関数である．数値属性の場合は値が小さいほど，名義値属性の場合は値が大きいほど類似度が高いということになる．

3. @cosme データセット

今回使用したデータセットは，@cosme(注1) のレビューデータである．@cosme は，株式会社アイスタイル(注2) が運営する日本最大級の化粧品レビューサイトであり，1999 年 12 月にサー

(注1)：http://www.cosme.net
(注2)：http://www.istyle.co.jp/

ビスが開始された．ユーザーのレビューを中心として，化粧品の情報提供，オリジナル商品の企画などが行われており，サイトの利用者は 20 代の女性が主であることが分かっている．このデータセットは，2012 年 11 月に@cosme をクロールして取得したものであり，131238 アイテム，480182 ユーザー，6606951 レビュー，20307 ブランドを有する．各レビューに含まれる情報は，ユーザー，アイテム，ブランド，得点，投稿時間，コメントであり，レビューの得点は 0〜7 の整数値をとりうる．また，得点を付与しないコメントだけのレビューも投稿可能である．

4. 結果と考察

今回は，プロット図の見易さを考慮して，レビュー回数が 300 以上の 1190 ユーザを分析対象とし，s を日単位に離散化させ，Δ =365 日，$K = 15$ とした．ユーザ属性類似度は，ユーザ v と他のユーザ全員 ($u \in V \setminus \{v_0, v\}$) との $AV_{u,v}$，$NV_{u,v}$ を算出し，それらを v が属する $\mathcal{C}_k$ ごとに平均値化したものを，$\overline{AV_{v,k}}$，$\overline{NV_{v,k}}$ とする．なお，$AV_{u,v}$ は年齢，$NV_{u,v}$ は肌質・髪質・髪量・血液型を値として扱い，$\boldsymbol{q}$ は，属性の次元数 $(6, 3, 3, 4)$ に応じて，$\boldsymbol{q} = (1.50, 0.75, 0.75, 1.00)^T$ に設定した．

図 1〜図 15 に $\mathcal{C}_1$〜$\mathcal{C}_{15}$ のプロット図を示す．赤い太線は代表オブジェクト $p_k \in \mathcal{P}$ に選定されたユーザで，青い線はそれ以外のユーザである．表 1 と表 2 に示す値は，$\mathcal{C}_k$ に属するユーザ群の $\overline{AV_{v,k}}$ と $\overline{NV_{v,k}}$ のさらに平均をとったものである．以下，実験結果についての考察を述べる．

$\mathcal{C}_1$，$\mathcal{C}_2$，$\mathcal{C}_4$，$\mathcal{C}_{15}$ に属するユーザは，トップユーザとの類似度に極端な変動があまり見られないため，古くから常駐しているユーザであるといえる．加えて，常駐ユーザを長期的な類似度の変化でさらに分けられることが見て取れる．$\mathcal{C}_3$ と $\mathcal{C}_{12}$ に属するユーザは，近年になるにつれてレビューの頻度が少なくなった先行類似型，いわば古参であることがうかがえる．この 2 つのクラスタに属する $\overline{AV_{v,k}}$ は高めであり，且つユーザの平均年齢も高いので，古くからユーザ登録をしていた可能性が高い．逆に，$\mathcal{C}_8$ と $\mathcal{C}_{10}$ は，$\overline{AV_{v,k}}$ が高め且つ平均年齢が低いので，後続類似型，いわば新参であることが示唆される．さらに，$\mathcal{C}_3$，$\mathcal{C}_{12}$，$\mathcal{C}_8$，$\mathcal{C}_{10}$ は，自クラスタでの $\overline{NV_{v,k}}$ が，他クラスタでのものより比較的低めなので，この分類には名義値属性の類似度があまり寄与していないことも分かる．一方，$\mathcal{C}_7$ に属するユーザは，ある期間だけレビューが活発になったイベント型と言えよう．$\mathcal{C}_7$ のユーザは $\overline{AV_{v,k}}$ が低く，平均年齢も低いので，かなり限定された若い年齢層によるバーストを検出したと見ることもできる．また，$\mathcal{C}_9$ の $\overline{NV_{v,k}}$ はどのクラスタ間でも低く，且つ全体で最も低いため，名義値属性を設定していないユーザ群による動きがあったことが想定できる．

5. ま と め

レビューサイトにおけるトップユーザとの動的レビュー類似度を定式化し，それを用いてユーザを K-median 法でクラスタリングすることを試みた．現実の大規模レビュー時系列データを用いた実験結果より，ユーザ間の類似度は，動的に多様な変化をすることが分かった．さらに，本分析手法は，視覚的にわかりやすい分類が可能であり，明確なユーザクラスタリングをする際の利用価値を示した．今後は，トップユーザに限らず，分析対象としたユーザ全員に対して手法を適応し，それらの結果をさらにクラスタリングすることを検討する．

謝辞 本研究は，株式会社豊田中央研究所との共同研究および，科学研究費補助金基盤研究 (C)(No.23500312) の補助を受けた．

文 献

[1] Salganik, M. J., Dodds, P. S. and Watts, D. J.: Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market, *Science*, Vol. 311, No. 5762, pp. 854–856 (2006).

[2] 木村，斉藤，中野，元田：社会ネットワークにおける有力ノード抽出のための情報拡散モデルの学習，人工知能学会誌，Vol. 25, pp. 215–223 (2010).

[3] Nemhauser, G. L., Wolsey, L. A. and Fisher, M. L.: An analysis of approximations for maximizing submodular set functions, *Mathematical Programming*, Vol. 14, pp. 265–294 (1978).

表 1 $\mathcal{C}_k$ ごとの $\overline{AV_{v,k}}$ の平均値

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	7.28	6.91	7.87	7.43	7.20	7.36	6.86	7.82	7.47	7.90	7.01	7.97	7.58	7.00	7.58
2	6.91	6.36	7.57	7.17	6.81	6.96	6.55	7.41	7.08	7.51	6.52	7.77	7.21	6.67	7.21
3	7.87	7.57	8.26	7.82	7.66	7.95	7.33	8.41	8.02	8.55	7.66	8.20	8.14	7.46	8.07
4	7.43	7.17	7.82	7.31	7.24	7.56	6.79	7.97	7.61	8.14	7.24	7.72	7.72	6.98	7.65
5	7.20	6.81	7.66	7.24	6.95	7.25	6.69	7.78	7.34	7.89	6.93	7.66	7.48	6.80	7.42
6	7.36	6.96	7.95	7.56	7.25	7.36	7.01	7.92	7.54	7.95	7.07	8.09	7.63	7.09	7.64
7	6.86	6.55	7.33	6.79	6.69	7.01	6.06	7.41	7.04	7.63	6.65	7.22	7.19	6.40	7.12
8	7.82	7.41	8.41	7.97	7.78	7.92	7.41	8.18	7.99	8.34	7.50	8.57	8.09	7.58	8.09
9	7.47	7.08	8.02	7.61	7.34	7.54	7.04	7.99	7.62	8.09	7.19	8.10	7.76	7.18	7.74
10	7.90	7.51	8.55	8.14	7.89	7.95	7.63	8.34	8.09	8.37	7.59	8.77	8.15	7.72	8.21
11	7.01	6.52	7.66	7.24	6.93	7.07	6.65	7.50	7.19	7.59	6.64	7.85	7.30	6.76	7.31
12	7.97	7.77	8.20	7.72	7.66	8.09	7.22	8.57	8.10	8.77	7.85	7.92	8.26	7.42	8.12
13	7.58	7.21	8.14	7.72	7.48	7.63	7.19	8.09	7.76	8.15	7.30	8.26	7.82	7.30	7.85
14	7.00	6.67	7.46	6.98	6.80	7.09	6.40	7.58	7.18	7.72	6.76	7.42	7.30	6.56	7.24
15	7.58	7.21	8.07	7.65	7.42	7.64	7.12	8.09	7.74	8.21	7.31	8.12	7.85	7.24	7.81

表 2 $\mathcal{C}_k$ ごとの $\overline{NV_{v,k}}$ の平均値

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.09	1.03	1.04	1.06	1.01	1.05	1.05	1.05	0.88	1.03	1.06	1.00	1.09	1.00	1.10
2	1.03	0.95	0.99	0.99	0.96	1.02	1.03	1.01	0.89	0.97	1.01	0.95	1.02	0.97	1.02
3	1.04	0.99	0.98	1.00	0.96	1.01	0.99	0.99	0.86	0.97	1.02	0.94	1.02	0.98	1.03
4	1.06	0.99	1.00	0.99	0.96	1.00	1.00	1.00	0.86	0.98	1.01	0.95	1.02	0.97	1.05
5	1.01	0.96	0.96	0.96	0.90	0.97	0.97	0.97	0.84	0.94	0.99	0.92	0.99	0.93	1.00
6	1.05	1.02	1.01	1.00	0.97	1.03	1.07	1.06	0.91	1.00	1.04	0.99	1.06	1.05	1.04
7	1.05	1.03	0.99	1.00	0.97	1.07	1.00	1.06	0.90	0.98	1.01	0.98	1.06	0.96	1.04
8	1.05	1.01	0.99	1.00	0.97	1.06	1.06	0.98	0.90	0.98	1.02	0.98	1.05	0.99	1.03
9	0.88	0.89	0.86	0.86	0.84	0.91	0.90	0.90	0.76	0.85	0.87	0.85	0.88	0.86	0.88
10	1.03	0.97	0.97	0.98	0.94	1.00	0.98	0.98	0.85	0.91	1.00	0.93	1.00	0.96	1.01
11	1.06	1.01	1.02	1.01	0.99	1.04	1.01	1.02	0.87	1.00	0.99	0.96	1.05	1.01	1.06
12	1.00	0.95	0.94	0.95	0.92	0.99	0.98	0.98	0.85	0.93	0.96	0.89	0.99	0.93	0.98
13	1.09	1.02	1.02	1.02	0.99	1.06	1.06	1.05	0.88	1.00	1.05	0.99	1.05	1.00	1.07
14	1.00	0.97	0.98	0.97	0.93	1.05	0.96	0.99	0.86	0.96	1.01	0.93	1.00	0.99	1.01
15	1.10	1.02	1.03	1.05	1.00	1.04	1.04	1.03	0.88	1.01	1.06	0.98	1.07	1.01	1.07

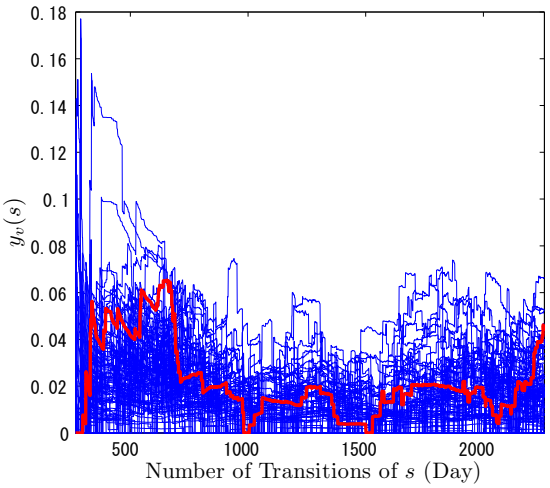

図 1　$\mathcal{C}_1$のプロット

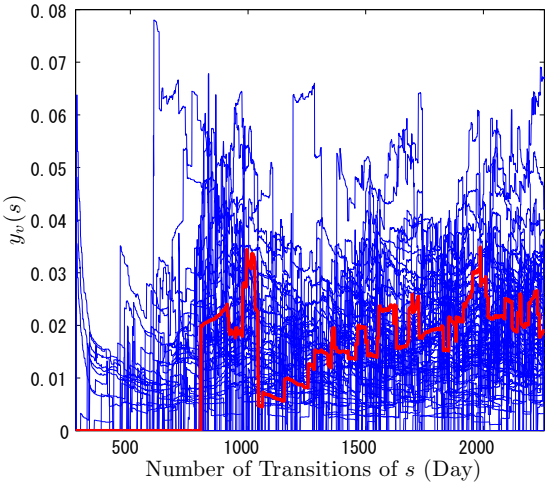

図 2　$\mathcal{C}_2$のプロット

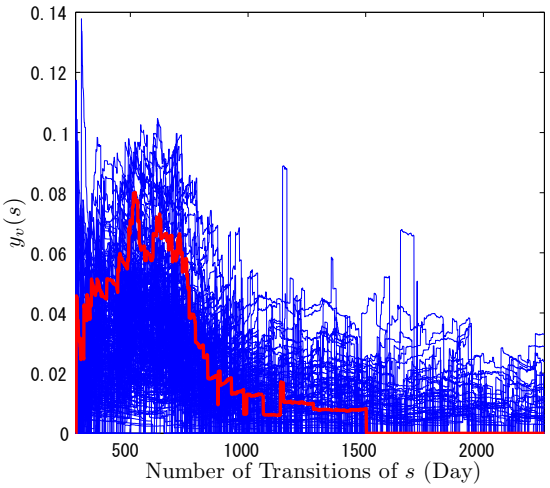

図 3　$\mathcal{C}_3$のプロット

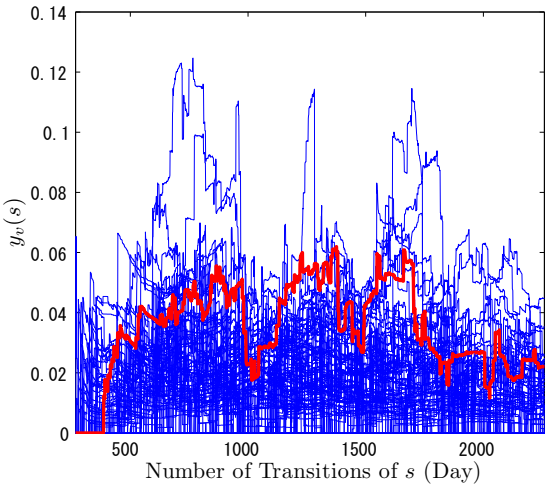

図 4　$\mathcal{C}_4$のプロット

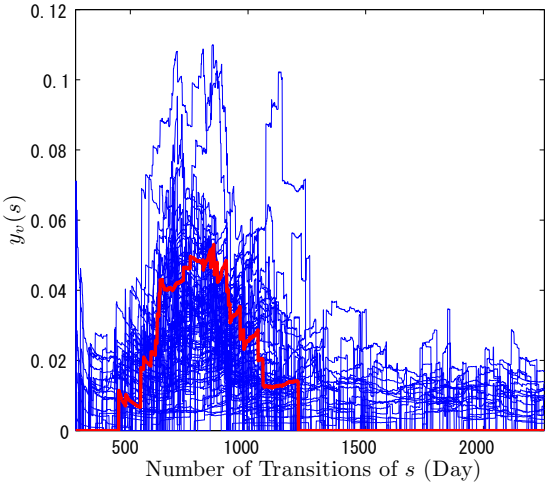

図 5　$\mathcal{C}_5$のプロット

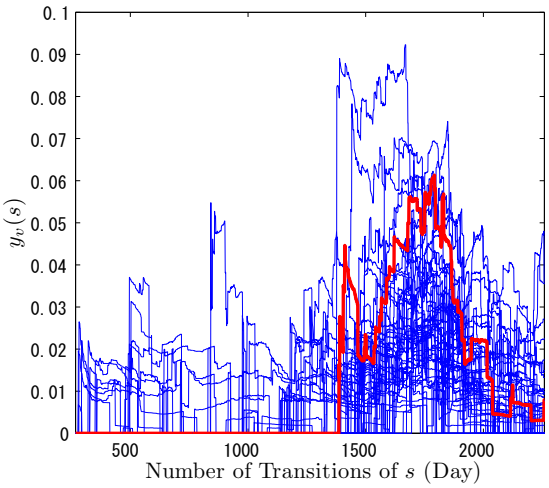

図 6　$\mathcal{C}_6$のプロット

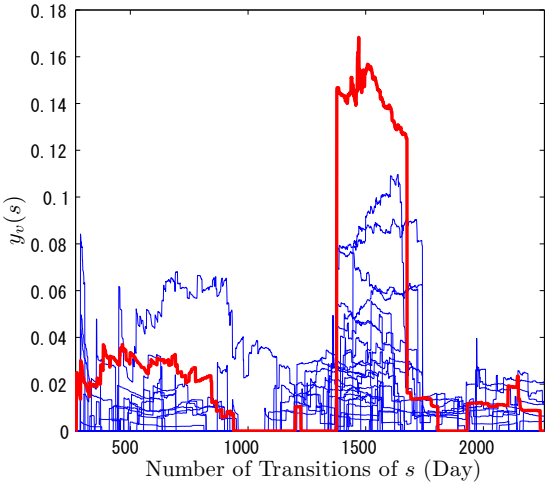

図 7　$\mathcal{C}_7$のプロット

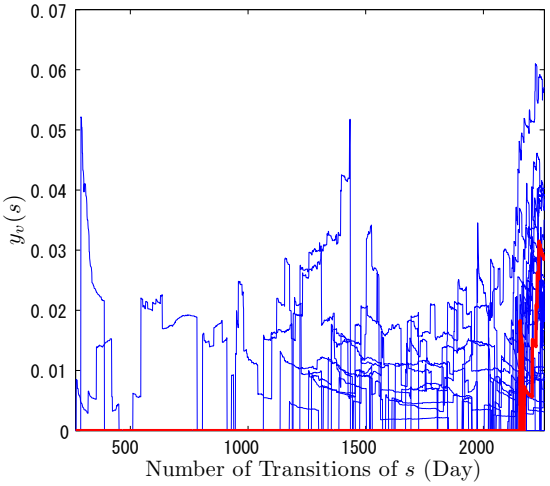

図 8　$\mathcal{C}_8$のプロット

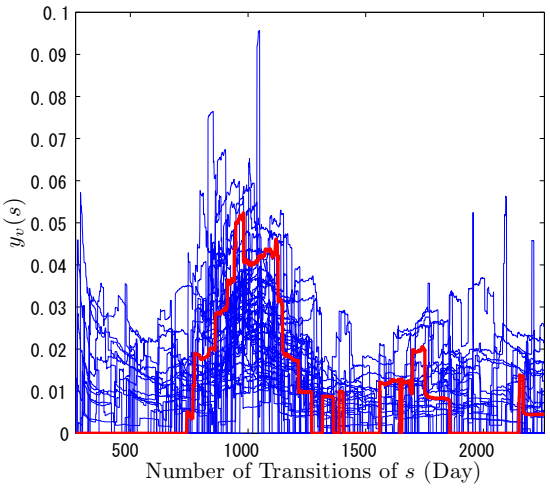

図 9　$\mathcal{C}_9$のプロット

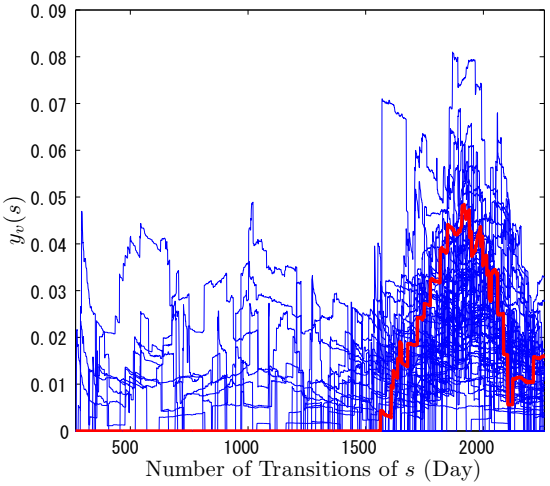

図 10　$\mathcal{C}_{10}$のプロット

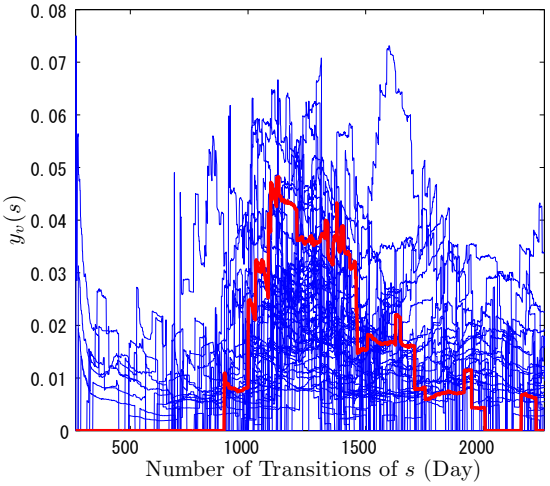

図 11　$\mathcal{C}_{11}$のプロット

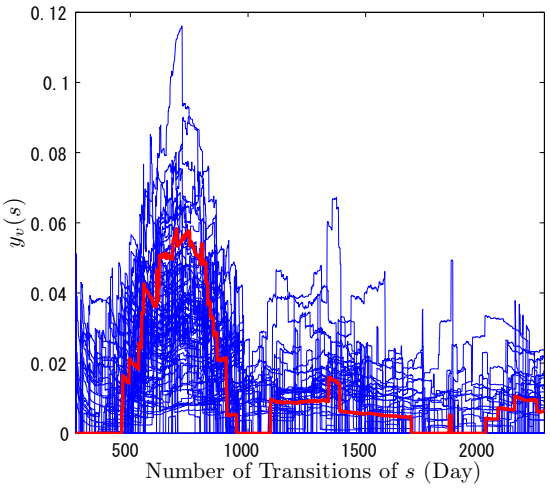

図 12　$\mathcal{C}_{12}$のプロット

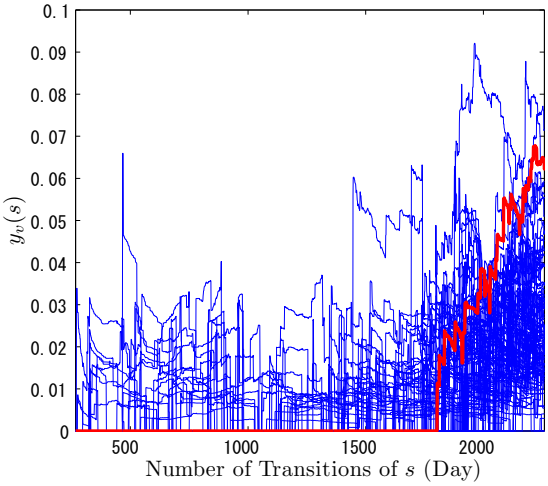

図 13　$\mathcal{C}_{13}$のプロット

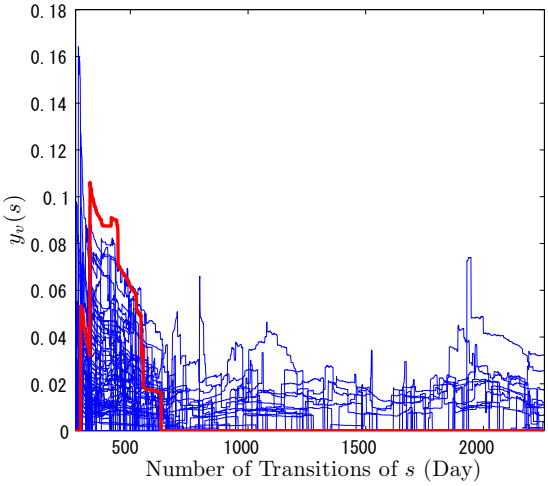

図 14　$\mathcal{C}_{14}$のプロット

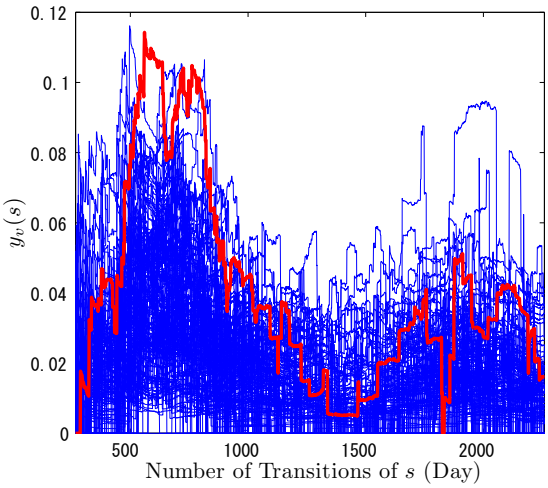

図 15　$\mathcal{C}_{15}$のプロット