ノード分類に基づくネットワーク上のシミュレーションにおける重要ノードの抽出法

湯浅友幸
東京大学大学院 工学系研究科
システム創成学専攻 白山研究室

概要

- 背景
- 目的
- 提案手法

–シミュレーションモデル
–特徴量によるノード分類
–分類に基づくシミュレーションの可視化
–ネットワークモデル

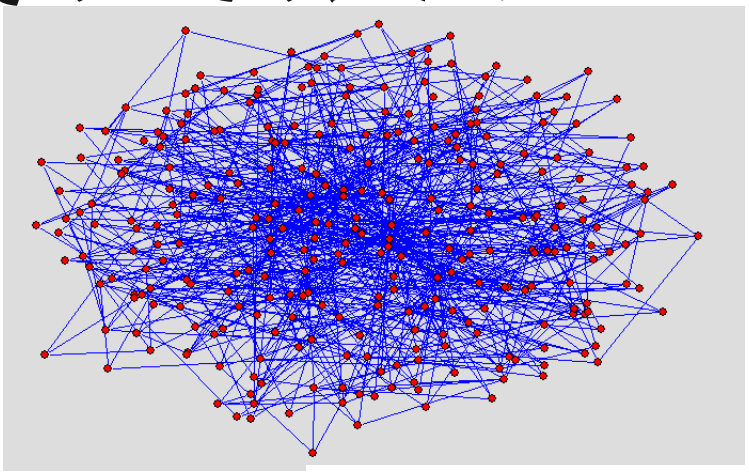
- 実験と考察
- 結論
- 付録
- 参考文献
- 謝辞

背景

- 感染症の流行や噂の伝播，集団の合意形成などの世の中の諸現象を背後に存在するネットワーク構造に基づき分析する研究が行われている[1]
- これらの研究の多くは人間をノード，人間どうしのつながりをリンクで模したネットワーク上で現象をモデル化したシミュレーションによるものである
- こうした研究の成果としてネットワーク全体に定義されるマクロな構造的特徴量（次数分布，次数相関など[1]）が現象に及ぼす影響が明らかにされてきた

背景

- 一方，よりミクロな視点から分析を行い，シミュレーションにおいて重要な役割を果たすノードを見つける試みはあまりなされていない
- 重要ノードの抽出にはシミュレーションの可視化が有効と考えられるが，右下図のようにネットワークの可視化自体が煩雑になるため，その上でのシミュレーション動向の分析は難しくなる



※図はpajek[2]で作成

背景

- 重要ノード抽出のもう1つの方向性としてはネットワーク内の各ノードが持つ特徴量（次数，媒介中心性など[1]）に基づき重要ノードを発見する方法がある
- しかし，既存研究ではデータ分析の困難さゆえに扱う特徴量が限定的であり，複数の特徴量の影響が考慮されていない（例えば，既存研究では次数が高いノードが重要な役割を果たす程度のことしかわかっていない）
- また，抽出された重要ノードを含めた各ノードがシミュレーションにおいてどんな役割を果たすかが詳しく分析されていない

目的

- 本研究では，ネットワーク上のシミュレーションにおける重要ノード抽出のための新たな手法を提案する
- 提案手法では，複数の特徴量が定義されるネットワーク内の各ノードをデータマイニングの手法を用いることにより効率的に分類する
- そして，分類に基づくシミュレーションの可視化を行うことにより各ノードのシミュレーションにおける役割と重要ノードの抽出を行う

提案手法(シミュレーションモデル)

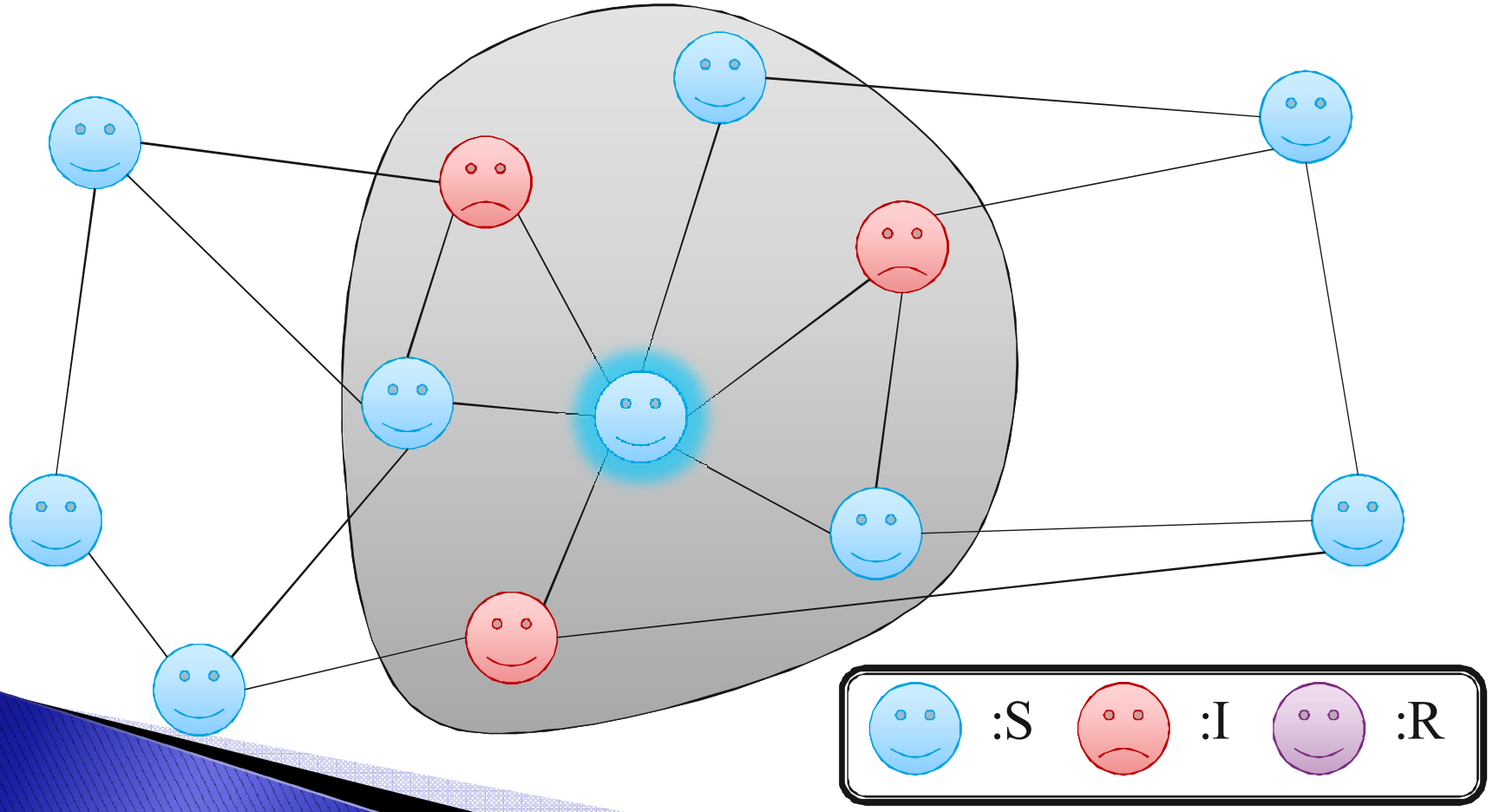
- ネットワーク上のシミュレーションモデルは多岐にわたるが，本研究では感染症の伝播を模したネットワーク上のSIRモデルを用いる
- このモデルでは人間をノード，人間どうしの接触をリンクとした人間の接触関係(友人・家族関係など)のネットワーク(コンタクト・ネットワーク)想定する
- そして，各ノードにS(Susceptible;感受性人口)，I(Infectious;感染人口)，R(Recovered;快復人口)の3属性を付与することで感染症の伝播をシミュレートする
- 次ページより具体的な手順を記す

提案手法（シミュレーションモデル）

- ネットワーク上のSIRモデル

・(a)~(c)のプロセスを繰り返すことで微少時間dtが進行

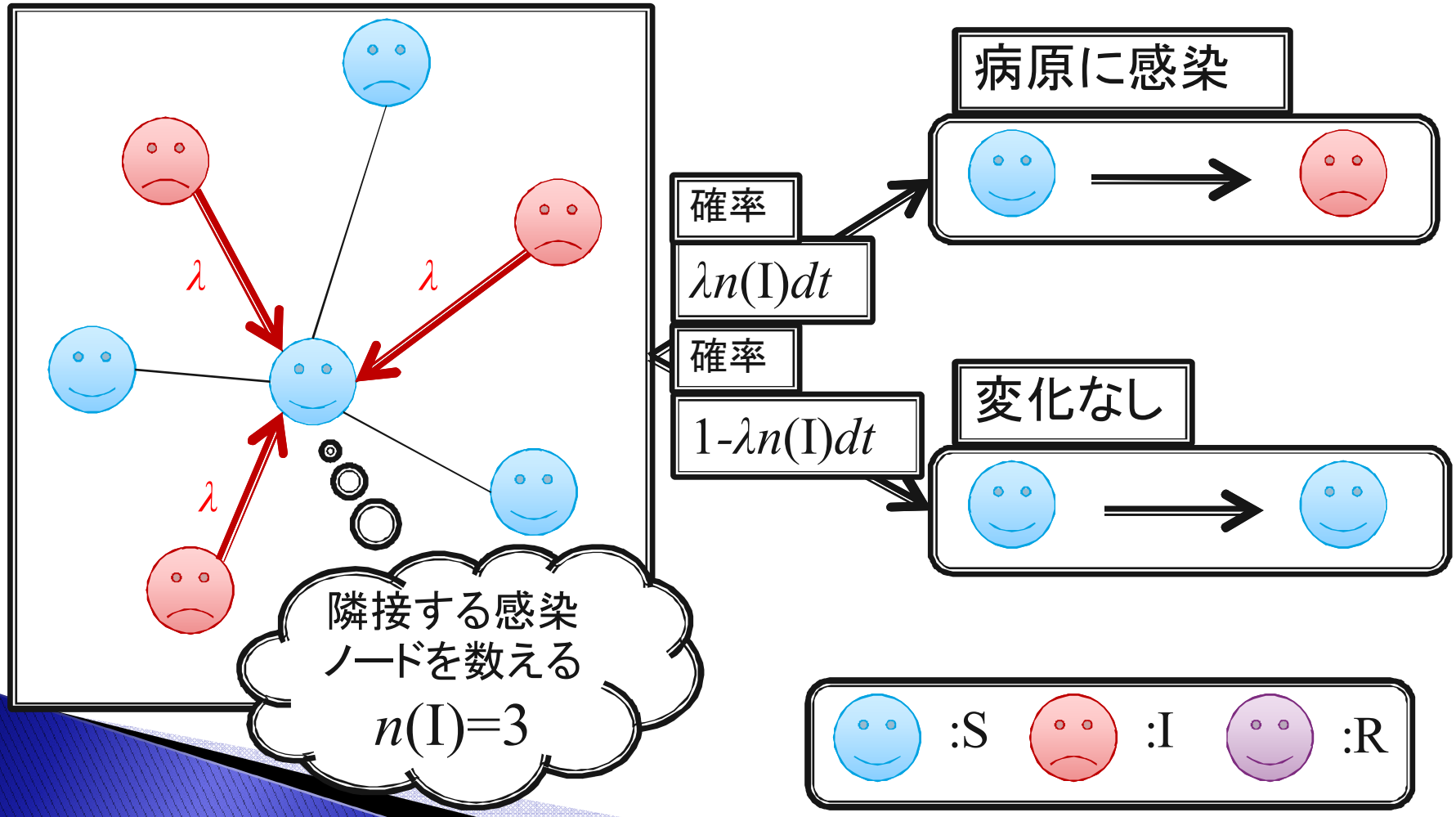
(a) ランダムにノード1つを選択し，これと隣接ノードに着目



提案手法(シミュレーションモデル)

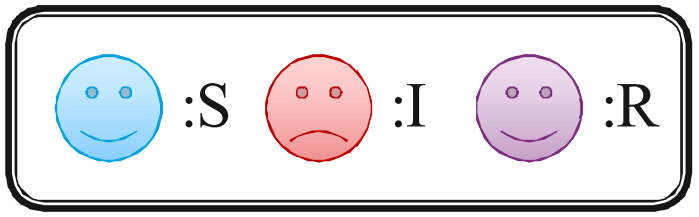
- ネットワーク上のSIRモデル

(b-1) 選ばれたノードが状態Sの場合(λは感染率)



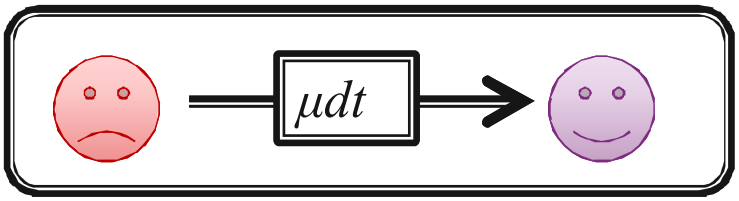
提案手法（シミュレーションモデル）

- ネットワーク上のSIRモデル



(b–2) 選ばれたノードが状態Iの場合

→治癒率μに比例する確率で快復



(b–3) 選ばれたノードが状態Rの場合

→変化なし

(c) (a), (b)をノード数の回数繰り返し，微小時間dtが進行

提案手法(特徴量によるノードの分類)

- 本研究では，ノード分類に用いる特徴量として次数，隣接ノードの平均次数，媒介中心性，平均頂点間距離，クラスター係数の5つを用いる
- 以降にその詳細を示す

＜次数k・隣接ノードの平均次数k_{nn}＞

–次数kとはノードの持つリンク数を示す

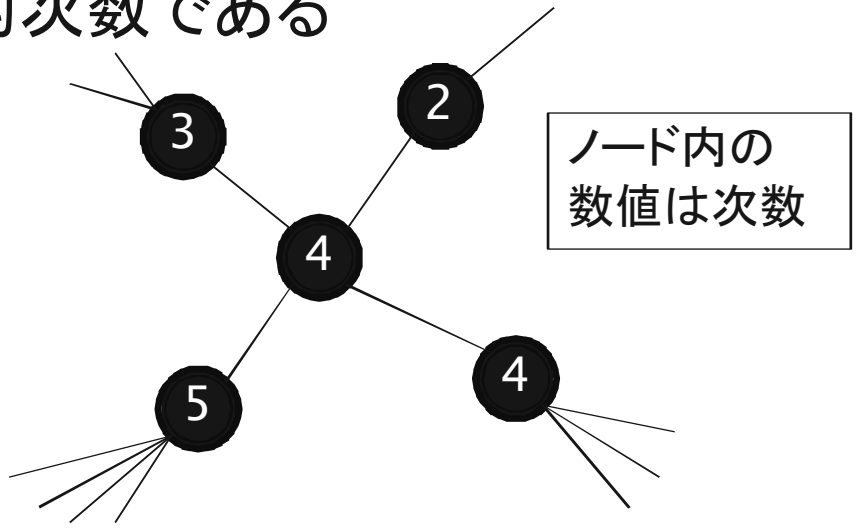
–k_{nn}はリンクするノードの平均次数である

–右図中央のノードでは，

$k=4$

$k_{nn}=(3+2+4+5)/4=3.5$

となる



提案手法(特徴量によるノードの分類)

＜媒介中心性b＞

–媒介中心性bはノードが他の任意の2ノード間をつなぐ最短路上に存在する割合を示す特徴量である

–bが大きいほどネットワーク上の情報の伝播などに大きな役割を果たすといわれている

–ノードiの媒介中心性b_iを数式で表すと以下になる

$$b_i \equiv \frac{\sum_{i_s=1;i_s\neq i}^{N}\sum_{i_t=1;i_t\neq i}^{i_s-1}\frac{g_i^{(i_s i_t)}}{N_{i_s i_t}}}{(N-1)(N-2)/2}$$

i_s ;始点ノードの番号

i_t ;終点ノードの番号

N ;ノード数

$g_i^{(i_s i_t)}$; i_sからi_tへのiを通る最短路数

$N_{i_s i_t}$;i_sからi_tへの最短路の総数

提案手法(特徴量によるノードの分類)

＜平均頂点間距離L＞

- 平均頂点間距離はノード間の距離の指標である
- ノードiの平均頂点間距離はiからネットワーク内の他の全ノードへの最短路の距離の平均として算出される
- ノードiの平均頂点間距離L_iを数式で表すと以下になる

$$L_i = \sum_{i \neq j} \frac{d(i,j)}{N-1}$$

$d(i,j)$;ノードi,j間の最短路の距離

N ;ノード数

提案手法(特徴量によるノードの分類)

＜クラスター係数C＞

- クラスター係数はノード周りのリンクの凝集度の指標である
- ノードiのクラスター係数はノードiの周りの三角形の個数に基づき定義される
- ノードiのクラスター係数C_iを数式で表すと以下になる

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

k_i ;ノードiの次数

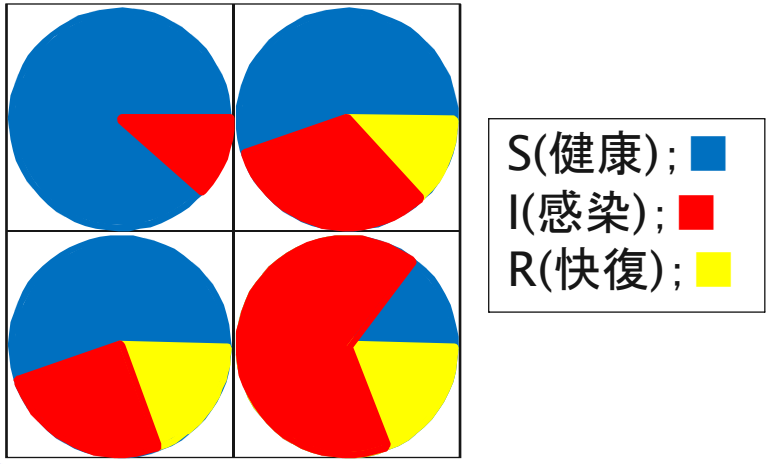
E_i ;ノードiの隣接ノード間に存在するリンク数
(ノードiを含む三角形の数)

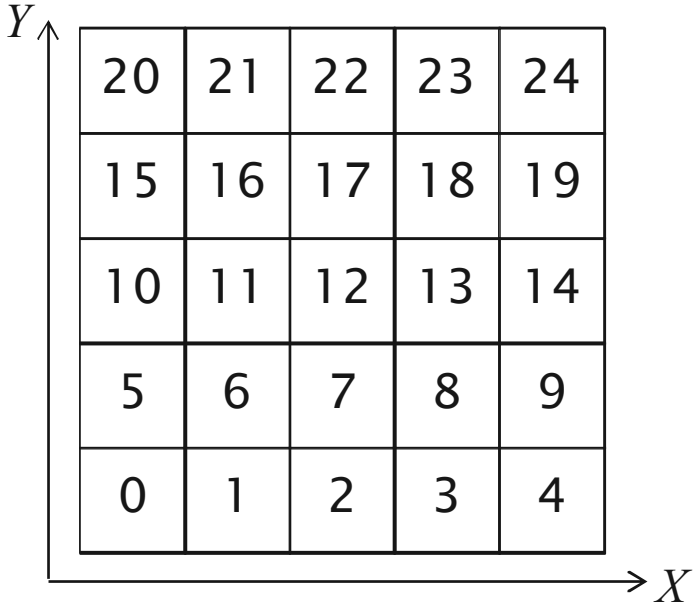
提案手法(特徴量によるノードの分類)

- 以上の5つの特徴量に基づき，ネットワーク内のノードを分類する
- 本研究ではVisual Mining Studioの自己組織化マップ(SOM)のツールを用いてノードの分類を行う
- 各ノードをk, k_{nn}, b, L, C に基づいて格子状(本研究では5×5とする)のSOM上に分類する
- 分類の傾向を各特徴量のカテゴリ内の平均値に対するヒートマップにより把握する
- SOMは近い特徴を持つノードを近いカテゴリに分類するため，後のシミュレーションの可視化において他のマイニング手法に比して現象の理解がしやすくなる

提案手法
（分類に基づくシミュレーションの可視化）

- シミュレーションの可視化
  - SOM上でシミュレーションを可視化する
  - 各カテゴリ内にS,I,R各状態のノード割合を示す円グラフを時系列に従って表示（左下図に例を示す）
  - また，各カテゴリに右下図のような番号を付ける

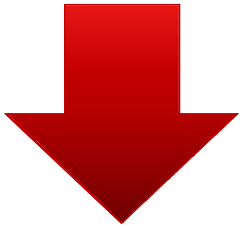




提案手法
（分類に基づくシミュレーションの可視化）

- 重要ノードの抽出

–SOMの分類傾向と可視化したシミュレーションの動向を照らし合わせることで，シミュレーションにおいて各カテゴリのノードがいかなる役割を果たすか，どのカテゴリのノードが重要な役割を果たすかを探る



ノード分類に基づいてネットワーク上の
シミュレーションにおける**重要ノード**を抽出

提案手法（ネットワークモデル）

- 本研究では，シミュレーションを行うネットワークとして VázquezのCNNモデル[3]を用いる
- CNNモデルは現実のコンタクト・ネットワークの，

① 非常に次数の高いハブが存在
→普通の人より圧倒的に多くの人に接触する人がいる
② 次数の高いノードどうしはリンクしやすい
→顔の広い人どうしは接触しやすい
③ クラスター係数が高くなりやすい
→友達の友達は友達になりやすい
という3つの特徴を再現するネットワークを生成するモデルである
※詳細な生成メカニズムは付録に示す

実験と考察

- 実験概要（ネットワーク生成とノードの分類）

–CNNモデルのメカニズムによって実験用のネットワークを生成する（本研究ではノード数を10000，平均次数を約8とする）

※ネットワークの生成は確率過程を含むため，毎回同じネットワークが生成されるとは限らない

⇒本研究で用いるネットワークは100回の予備実験により生成したネットワークの平均的な構造と有意な差がないことを確認している

–ネットワーク内のノードを提案手法に基づき分類する

実験と考察

- 実験概要（シミュレーションの実行と分析）

–生成ネットワーク上でSIRモデルのシミュレーションを行う

–初期感染ノードをランダムに10ノード配置し，系から状態Iのノードが消滅するまでシミュレーションを行う

–感染率λ=0.2，治癒率μ=1，離散化時間dt=0.01とする

※SIRモデルは確率過程を含むため常に同じ結果が出るとは限らない

⇒本研究で紹介するデータは100回の予備実験の平均的傾向と有意な差がないことを確認している

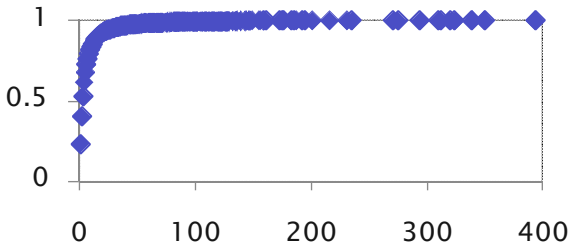
–提案手法に基づきシミュレーションを可視化，分析する

実験と考察

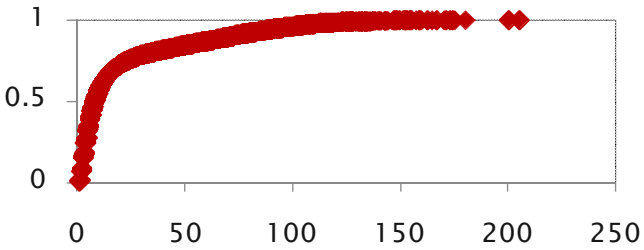
- ネットワークの生成

–ネットワーク内のノードの各特徴量に対する累積確率分布を示す

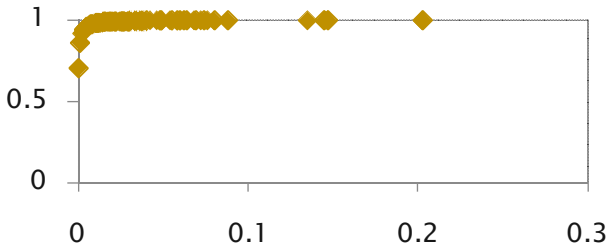
次数 k



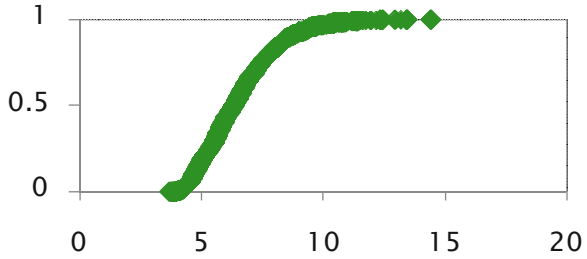
隣接ノードの平均次数 k_{nn}



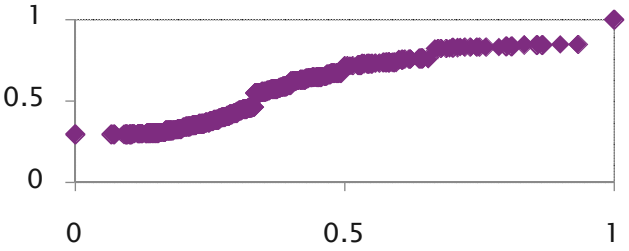
媒介中心性 b



平均頂点間距離 L



クラスター係数 C



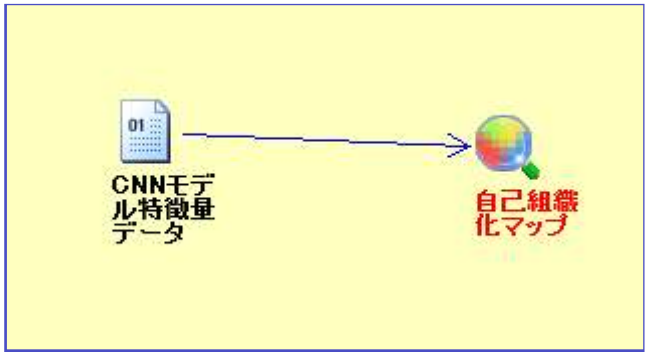
特徴量Xの累積確率分布
→縦軸はネットワーク内のノードの特徴量Xが横軸の値以下である割合を表す

実験と考察

- ノードの分類
  - Visual Mining StudioのSOMによりノードを分類する

number	k	knn	b	l	a
0	294	65.316	0.058	3.857	0.104
1	350	64.626	0.068	3.768	0.095
2	310	67.5	0.066	3.759	0.103
3	320	65.778	0.072	3.798	0.1
4	394	57.792	0.203	3.68	0.072
5	312	63.343	0.147	3.799	0.094
6	275	66.462	0.088	3.802	0.106
7	339	65.192	0.145	3.735	0.093
8	147	72.054	0.049	4.029	0.14
9	324	63.54	0.074	3.858	0.096
10	216	70.449	0.027	3.949	0.135
11	192	70.052	0.024	4.063	0.141
12	176	78.438	0.06	3.911	0.148
13	231	64.706	0.054	3.867	0.101
14	116	86.06	0.026	4.056	0.185
15	84	37.214	0.037	4.543	0.124
16	201	72.522	0.031	3.92	0.125
17	183	66.432	0.062	3.969	0.118
18	159	70.333	0.019	4.12	0.131
19	173	66.63	0.02	4.14	0.131
20	144	63.431	0.016	4.173	0.118
21	110	79.836	0.012	4.194	0.185
22	155	72.903	0.012	4.089	0.143
23	50	44.22	0.003	4.569	0.217
24	235	57.591	0.076	3.959	0.093
25	159	80.346	0.014	4.069	0.172

ノードの特徴量データ

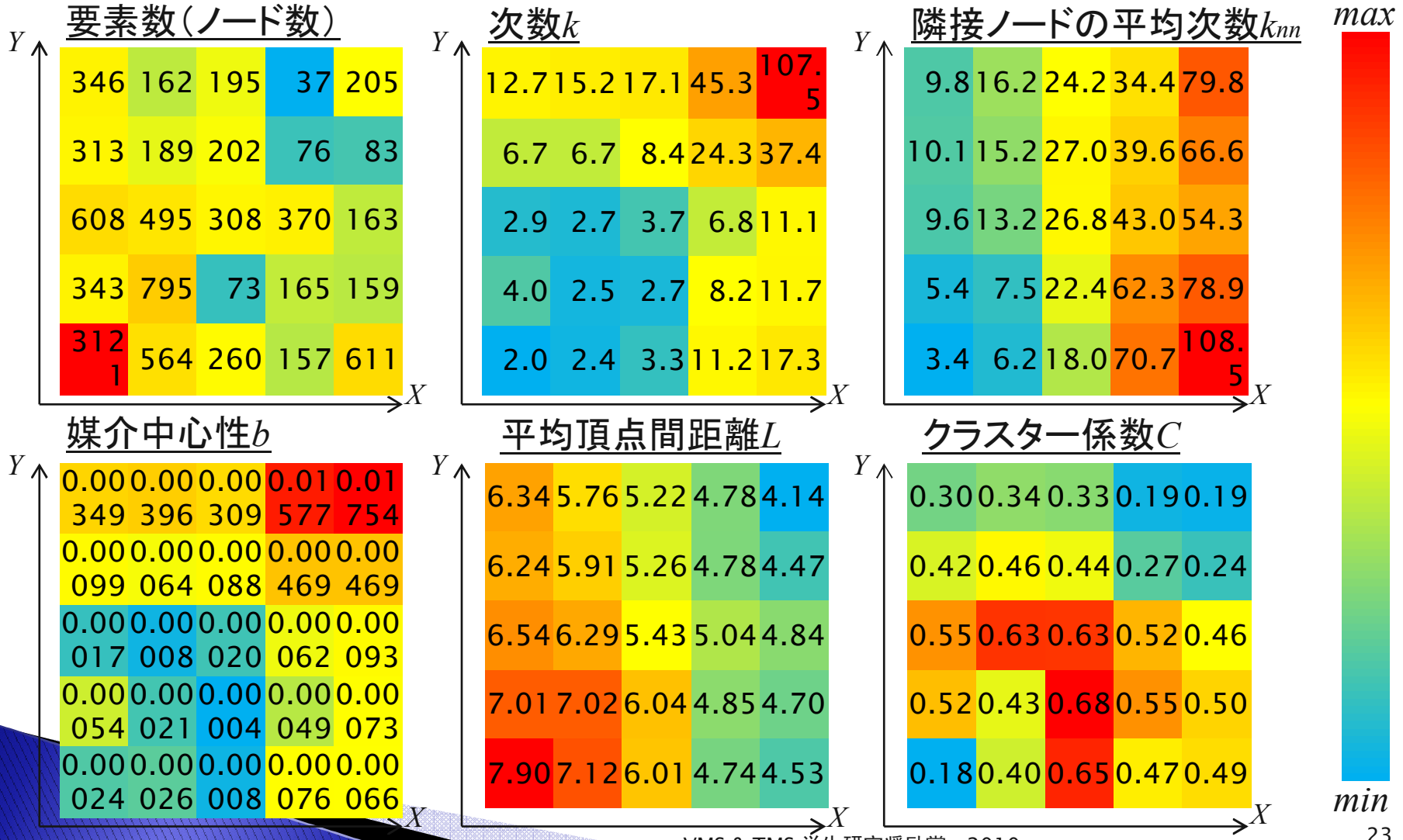


10000のノードを
25カテゴリに分類

※VMStudioのSOMの設定については付録に示す

実験と考察（SOMによる分類の結果）

- 各カテゴリ内のノードの要素数，特徴量の平均値に対するヒートマップを示す（各セルがカテゴリ，数値は特徴量の値）



実験と考察

- SOMによる分類の傾向

① 次数k

–より右側，上側のカテゴリほど高い傾向にある

② 隣接ノードの平均次数k_{nn}

–より右側のカテゴリほど高い傾向にある

–左側ではより上側の方が，右側ではより下側のカテゴリほど高い傾向にある

③ 媒介中心性b

–より右側，上側のカテゴリほど高い傾向にある

実験と考察

- SOMによる分類の傾向

④ 平均頂点間距離L

–より左側，下側のカテゴリほど高い傾向にある

⑤ クラスター係数C

–中央付近のカテゴリで高い傾向にある

–特に右上側のカテゴリで低い傾向にある

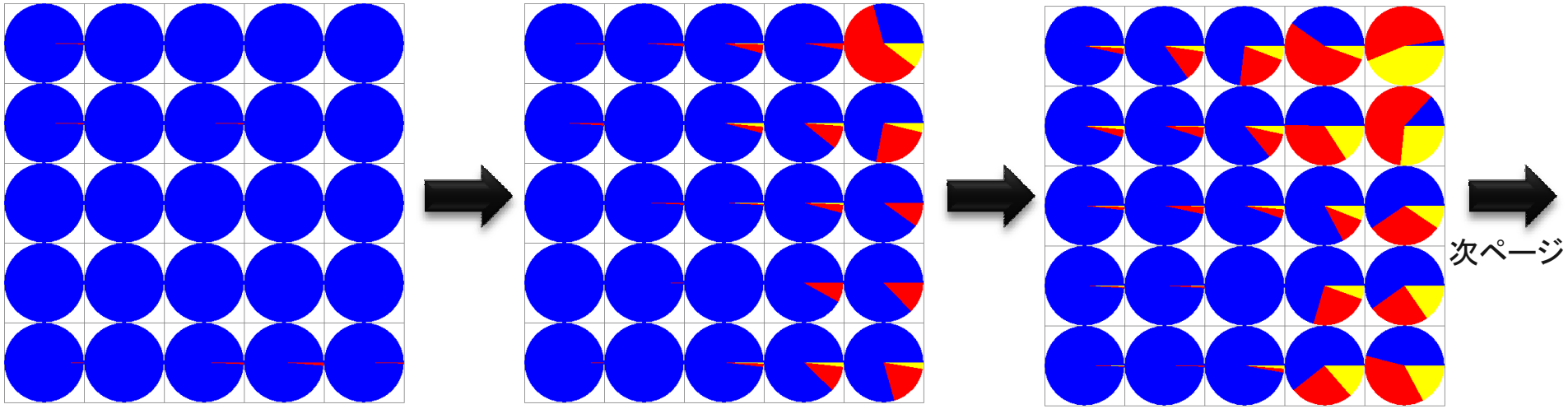
実験と考察

■;S(健康) ■;I(感染) ■;R(快復)

- シミュレーションの可視化

–時系列に従ってSOM上でシミュレーションを可視化する

t=0(初期状態)　t=0.5　t=1.0



初期状態:全10000のノードのうち10ノードが感染している

シミュレーション開始直後:右側のカテゴリを中心に感染開始
特にカテゴリ24で大規模に感染
4,19でも比較的多く感染

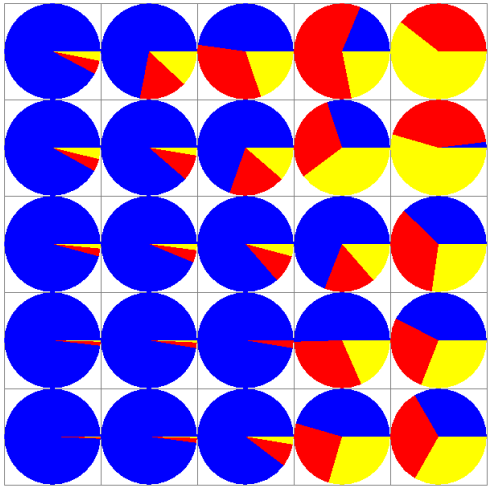
カテゴリ24にてほぼ全ノードが感染
23で一挙に感染拡大
右側のカテゴリで感染が進む
21,22で感染開始

実験と考察

■ ; S(健康) ■ ; I(感染) ■ ; R(快復)

- シミュレーションの可視化

t=1.5

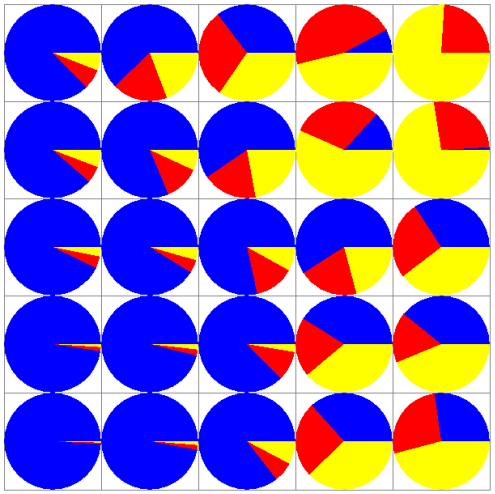


右側～中央のカテゴリで感染がさらに拡大

左側のカテゴリで感染が始まる

左下ではほとんど広がらない

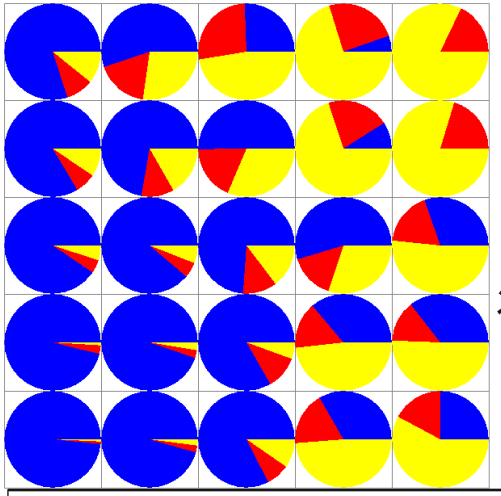
t=2.0



右上のカテゴリでは感染が拡大が止まり，快復のみの状態遷移になりつつある

左上のカテゴリでは一定の感染者割合を保って感染が拡大，同時に快復

t=2.5



右下のカテゴリでも感染の拡大が止まり，快復のみの状態遷移になりつつある

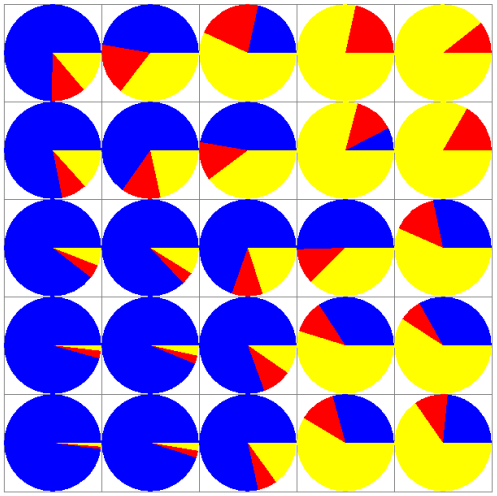
次ページ

実験と考察

■ ; S(健康) ■ ; I(感染) ■ ; R(快復)

- シミュレーションの可視化

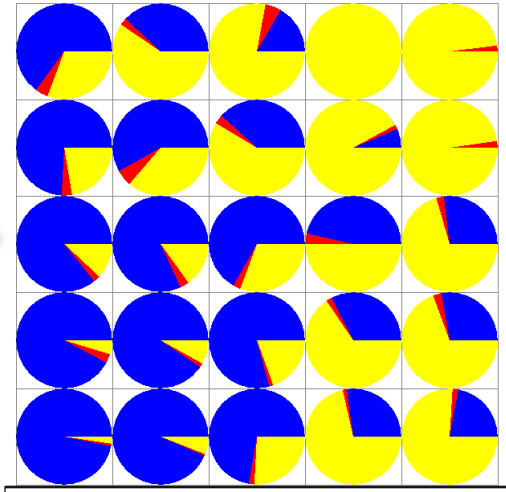
t=3.0



右側のカテゴリではほぼ快復のみになる

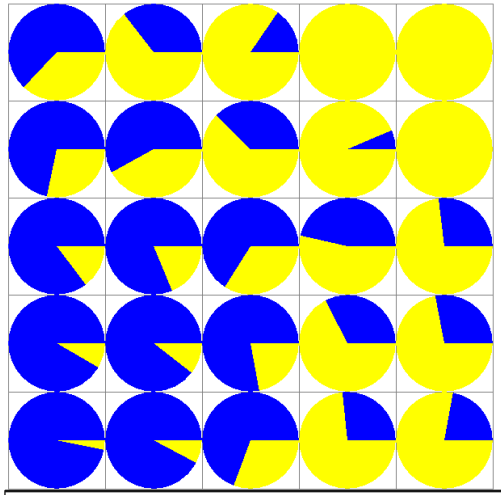
左上のカテゴリでは感染の拡大が止まり，快復のみの状態遷移になりつつある

t=5.0



すべてのカテゴリでほぼ快復のみがなされるようになる

t=12.6（終状態）



終状態：すべての感染ノードが系から消滅する

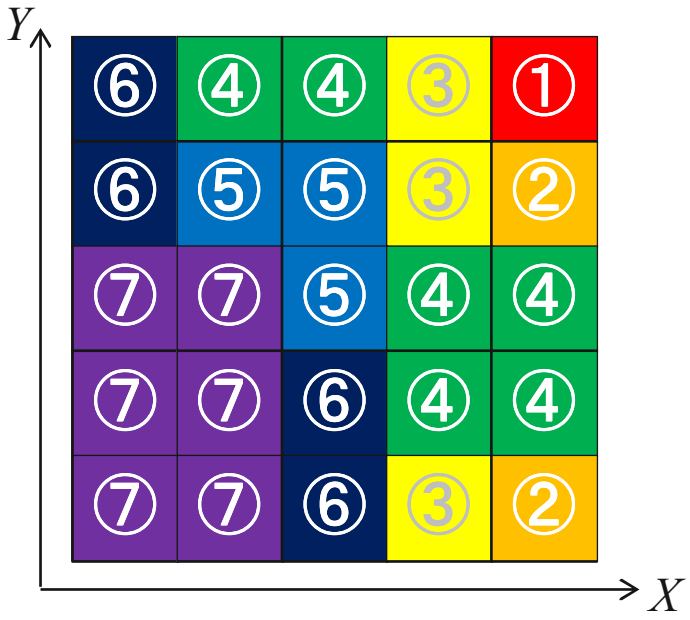
実験と考察

- SOMの分類とシミュレーション可視化に基づく考察

–シミュレーションにおけるカテゴリごとの感染拡大の順序を右下図に大まかに示す

–感染の拡大が早いカテゴリは前ページの終状態の感染状況から感染ノードの割合が多くなるカテゴリと一致することがわかる

–また，この順序は23ページの平均頂点間距離Lの小さな順にほぼ一致する



実験と考察

- SOMの分類とシミュレーション可視化に基づく考察

–一方，次数kと隣接ノードの平均次数k_{nn}の観点から結果を見ると，まず，

① 最初に感染するカテゴリ24はkが最大

② 次に感染するカテゴリ4はk_{nn}が最大

カテゴリ19はk,k_{nn}ともに高い

ことがわかる

–さらに，3番目に感染するカテゴリ群に関してはk,k_{nn}の両者が大きい，または，k_{nn}が非常に大きいことがわかる

実験と考察

- SOMの分類とシミュレーション可視化に基づく考察

–4,5,6番目に感染拡大するカテゴリにいついてはkまたはk_{nn}の値が大きなものから順次感染している

–ほとんど感染が拡大しなかった7番目の領域はk,k_{nn}とも非常に小さなカテゴリである

–媒介中心性bに関しては大きなカテゴリほど感染拡大が早い傾向も見受けられるが、必ずしもそれが成り立つわけではない

–クラスター係数Cに関しては小さいカテゴリほど感染が早い傾向も見受けられるが必ずしもそれが成り立つわけではない

–これらの結果に基づきシミュレーションにおける各ノードの役割を特に有意な差をもたらしたL,k,k_{nn}の3特徴量によってまとめる

実験と考察

- 各ノードの役割についての考察

–ネットワーク上の感染症伝播においてはLの小さなノードに始まり大きなノードに向けて感染が広がる
（Lはノードの感染タイミングを計る指標になりうる）

–一方，シミュレーションにおける各ノードの役割はk,k_{nn}によって以下のように特徴づけられる

① はじめに，最大のkを持つノード群が感染する

② 次に，このノード群がk_{nn}の大きなものを感染させる

③ さらに，これらのノード群がk,k_{nn}が中位のものを感染させ，k,k_{nn}が小さなものを最後に感染させる

※ ただし，k,k_{nn}が小さなノード群の感染は微少であり，これらのノードへの感染をもって感染は収束に向かう

実験と考察

- 重要ノードの抽出

–以上の考察から本シミュレーションにおいては

① 初期の感染拡大をもたらし，最大のkをもつカテゴリ24のノード

② カテゴリ24からの感染をk,k_{nn}が中位のノード群に橋渡しするカテゴリ4や19のノード（k_{nn}が特に大きく，kもカテゴリ24に準じて大きい）

③ 感染の収束をもたらすカテゴリ0,1,5,6,11,12のノード（k,k_{nn}ともに非常に小さい）

がとりわけ重要な役割を果たすノードといえる

実験と考察

- 知見の新規性

–第1に，既存研究[4]などにおいてはCNNモデルのような次数kの高いハブの存在するネットワークではハブからの感染拡大のみがシミュレーションの重要な要素と考えられていた

–一方で，本研究は隣接ノードの平均次数k_{nn}の高いノードもまた重要な役割を果たすことを示した

–第2に，kが非常に小さなノードの存在が感染を抑制するという知見は本研究の新たな知見である

–最後に，既存研究[4]などではkの大きなノードほど早く感染しやすいと述べているが，本研究では感染のタイミングの早さはkよりもむしろLに顕著に現れること，必ずしもkの大きなノードが先に感染するわけでないことを示している

結論と展望

- 本研究では，
  - ネットワーク上のシミュレーションにおいて重要な役割を果たすノードを抽出されていないという問題を解決するために，SOMによるノードの分類と可視化を用いた重要ノード抽出法を提案した
  - シミュレーションの一例としてネットワーク上のSIRモデルを扱い，人間の接触関係のネットワークの特徴をよく表すCNNモデル上でシミュレーションを実行した
  - 提案手法によりノードを分類し，シミュレーションの可視化結果とともに分析を行うことで，重要ノードの抽出を行い，既存研究で得られていなかった複数の新たな知見を得た

結論と展望

- 本研究の成果は，
  - 通常は困難なネットワーク上のシミュレーションの可視化を，データマイニングによるノード分類によって簡便化することによって得られた
  - これはデータマイニングの手法がデータ分析のみでなくシミュレーションの時系列的な可視化にも有効なことを示している

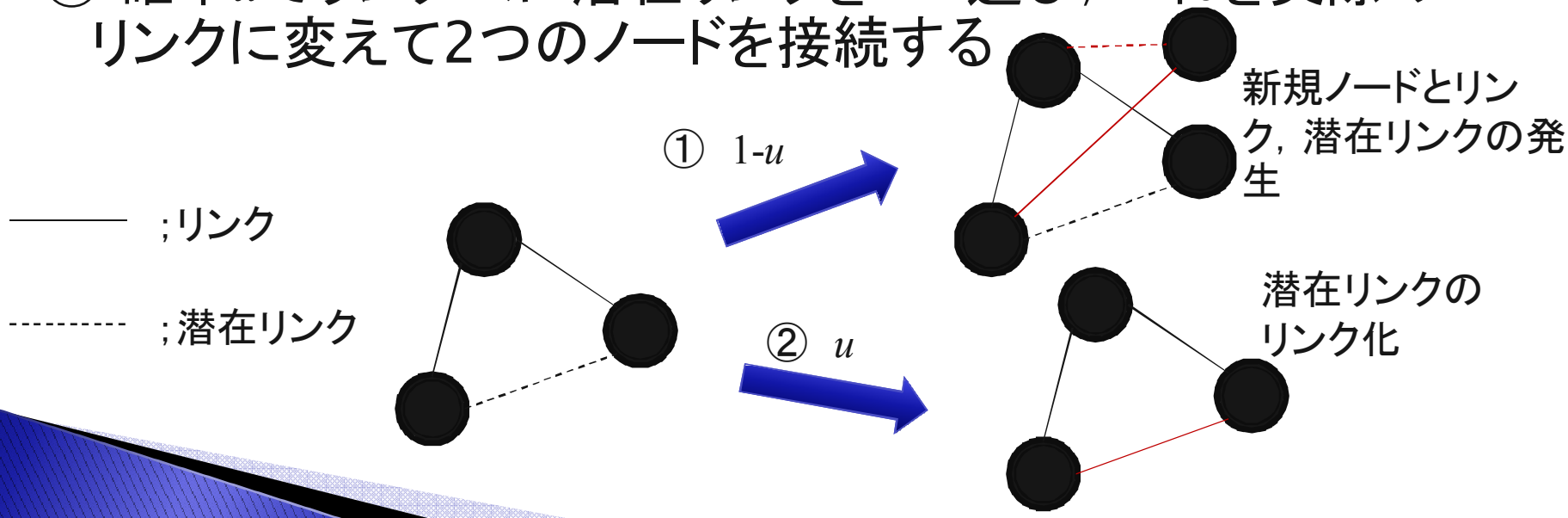
- 今後の展望としては，
  - 本手法をネットワーク上の他のシミュレーションにも適用し，その汎用性を示すことが考えられる

付録1(CNNモデルの生成法)

- CNNモデルでは確率u(本研究ではu=0.25)に基づいて以下の2つのプロセスを所望のノード数になるまで繰り返す

① 確率1-uで新規ノードをネットワークに追加し，ランダムに選んだ既存のノードvとリンクでつなぐ
また，vとリンクするノードと新規ノードを潜在リンクでつなぐ

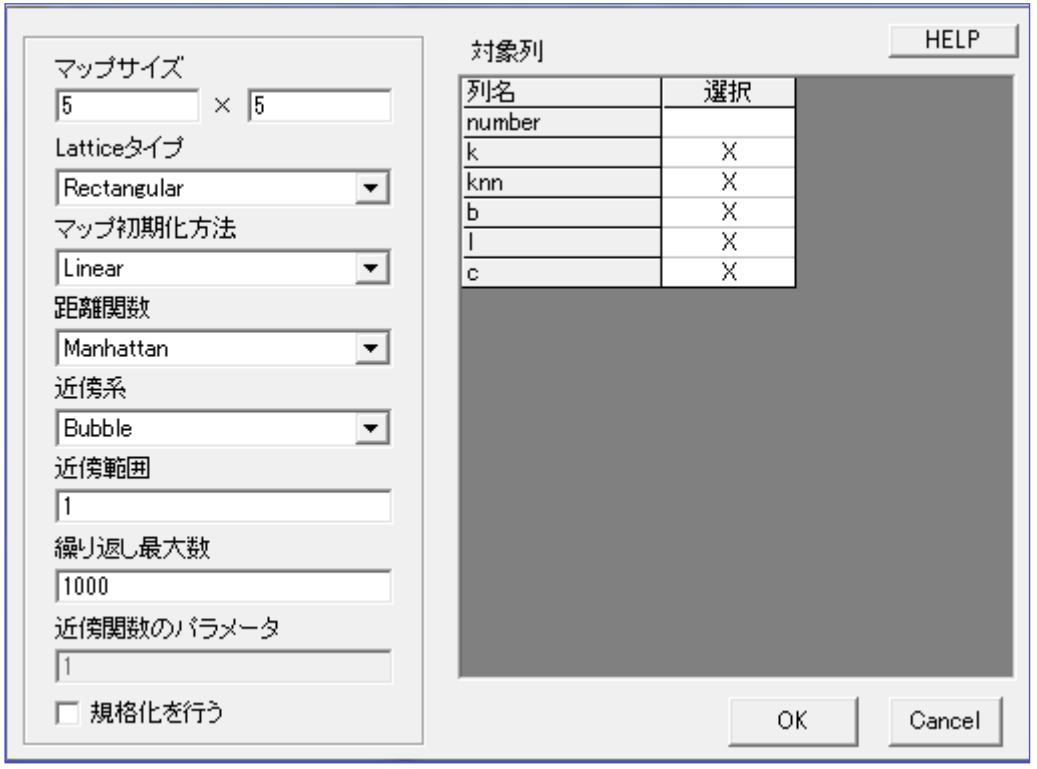
② 確率uでランダムに潜在リンクを1つ選び，これを実際のリンクに変えて2つのノードを接続する



付録2（VMStudio SOMの設定）

- VMStudioのSOMの設定は右図の通り
  - マップサイズ；5×5
  - Latticeタイプ；Rectangular
  - マップ初期化法；Linear
  - 距離関数；Manhattan
  - 近傍系；Bubble
  - 近傍範囲；1
  - 繰り返し回数；1000

である



参考文献

[1] 増田直紀, 今野紀雄. 複雑ネットワーク 基礎から応用まで. 近代科学者, 2010.

[2] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[3] A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(056104), 2003.

[4] Y.Moreno, R.Pastor-Satorras, and A.Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B*, 26(4):521-529, 2002.

謝辞

- 本研究を行うにあたりVisual Mining Studio，Text Mining Studioをご貸与いただいた株式会社数理システム様に心より御礼申し上げます