フリーソフトウェアのソーシャルライフ

——データ解析環境Rの拡張パッケージ——

荒　木　孝　治

1．はじめに

自然科学，工学，社会科学，人文科学等の様々な分野を対象としてネットワーク分析の手法が適用されている[1]。それは，こうした分野においてネットワーク構造を考えることができ，その構造を分析することにより，そこに内在するダイナミズムを知ることができると期待されているからである。近年では，オープンソースソフトウェアの世界を対象とするネットワーク分析も盛んになってきた［4, 5, 7, 13］。さらに，ウィキペディアという誰もが自由に執筆・編集できる百科事典の項目間の構造を分析したり［3, 14］，CVS（Concurrent Versions System）リポジトリに蓄えられている情報を分析したりと［6］，ネットワーク分析の手法の適用領域が拡大してきている。

本稿では，フリーソフトウェアRに注目し，そこに内在するネットワーク構造の分析を試みる。Rは，統計学はもとより，コンピュータ科学，バイオインフォマティクス，金融・マーケティングといった様々な分野における世界の研究者が協力して開発しているオープンソースのデータ解析環境である［9］。Rは，基本パッケージ（base package）と拡張パッケージ群（contributed packages）から構成される。基本パッケージは，コアチーム（R Development Core Team）と呼ばれる少数のメンバーから構成される開発者グループにより保守・開発が行われている。これに対して拡張パッケージは，様々な人たちがボランタリに開発し，フリーで公開しているユーティリティソフトウェアや様々な分野のデータ分析に特化した応用ソフトウェア等である。

Rに新しい機能を付加するとき，2つの方法がある。1つは，基本パッケージのみに依存する新しいパッケージを開発する方法であり，もう1つは，すでに存在する拡張パッケージ中の関数を利用しながら開発する方法である。「車輪の再発明」という言葉があるように，ソフトウェア開発の基本は，既存の確立されたライブラリ等を有効に利用することにあり，これにより開発期間を短縮したり，バグを回避したりすることが可能となる。既存のパッケージを利用して新しいパッケージを開発すると，パッケージ間の依存関係が生じ，ネットワークが構成さ

れる。本稿では，Rのパッケージが生み出しているこうしたネットワークの特徴を調べる。

2．データと方法

2.1　データとグラフ表現

データは，The Comprehensive R Archive Network（CRAN）[1]のミラーサイトの1つである東京大学より2006年7月25日に取得した[2]。この時点においてパッケージは792本あった。

様々なものごとの関係を表現するとき，グラフがよく用いられる。グラフとは，点（ノード）と，ノード間を結合する線（エッジまたは矢印）の集まりである。グラフは隣接行列Xを用いて数学的に記述することができる。今の場合，ノード数をnとするとき，Xの要素$x_{ij}(i=1,...,n, j=1,...,n)$を，

$$x_{ij}=\begin{cases}1 & \text{パッケージが}a_i\text{が}a_j\text{に依存しているとき}\\ 0 & \text{依存していないとき}\end{cases}$$

と定めればよい。

Rにおいて，パッケージの内容は，パッケージの基本的な情報を記述するファイル（DESCRIPTIONファイル）の中で，統一的に説明されている。例として，パッケージaspaceの内容を図1に示す。図1には上方より，パッケージ名，タイトル，バージョン，公開日，著者，管理者，説明，ライセンス形態等が記載されている。パッケージの依存関係は，Depends欄の

Depends: R (>= 2.0.1), adehabitat, ade4, gpclib

よりわかる。これは，パッケージaspaceは4つのパッケージ，R (>= 2.0.1)，adehabitat, ade4, gpclibに依存していることを意味する。「R (>= 2.0.1)」は，バージョン2.0.1以上のR本体に依存する（R本体で利用できる）ことを意味する。あらゆるパッケージはR本体に依存しているので，本稿ではこの依存関係を分析から除外する。するとaspaceは，adehabitat, ade4, gpclibという3つのパッケージに依存していることになる。この関係をグラフで表現すると，図2のようになる[3]。

1）CRAN：http://cran.r-project.org/

2）Rの拡張パッケージには大きく分けて，CRANで公開されているものとBioConductorという別の組織のサイト（http://www.bioconductor.org/）で公開されているものの2つがある。本稿ではCRANのもののみを分析対象としている。なお，BioConductorは，ゲノム解析に特化したオープンソースソフトウェアの開発プロジェクトである。

3）以下，依存関係のグラフの作成には，フリーソフトウェアのNetDraw［2］およびRを利用した。また，グラフに関する様々な計算および処理には，NetDrawおよびR，Rのパッケージ群を利用した。なお，Rのための拡張パッケージとしてpkgDepToolsの公開が，BioConductorのデベロッパー用メーリングリストで2006年9月5日にアナウンスされた。本稿の作成では利用していないが，これを用いると，依存関係のデータの取得およびネットワークデータへの変換が簡単にできるようである。

図2では，パッケージをノードで，依存関係を矢印で表している。ノードにはパッケージ名をラベルとして付けている。依存には「依存する・される」という関係があるため，グラフは有向グラフとなる。なお，グラフにおいてノードの位置や矢印の長さには基本的に意味はなく，依存関係のみが意味を持つ。

```
Package: aspace↓
Type: Package↓
Title: A collection of functions for estimating centrographic statistcs
and computational geometries from spatial point patterns↓
Version: 0.1↓
Date: 2006-05-25↓
Author: Tarmo K. Remmel, Ron N. Buliung↓
Maintainer: Ron N. Buliung <ron.buliung@utoronto.ca>↓
Description: A collection of functions for computing centrographic satis
tics (e.g., standard distance, standard deviation ellipse), and minimum
convex polygons (MCP)for observations taken at point locations.  A tool
is also provided for converting geometric objects associated with the ce
ntrographic statistics, and MCPs into ESRI Shapefiles.↓
License: GPL (Version 2 or later)↓
Depends: R (>= 2.0.1), adehabitat, ade4, gpclib↓
LazyData: yes↓
Packaged: Mon May 29 11:35:41 2006; remmelt↓
```

図 1　パッケージaspaceのDESCRIPTIONファイルの内容

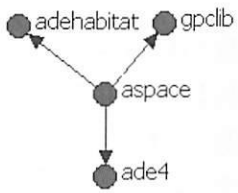

図 2　aspaceの依存関係

2.2　ネットワーク分析

ネットワーク分析において，ネットワークを特徴づけるための指標が様々提案されている。ここではWasserman and Faust［12］に基づいて，後で利用する代表的なものを概説する。

（1）中心性

中心性は，ノードがネットワークのなかで果たしている位置や役割を数値化する指標である。中心性の中で基本となるものが次数（degree）である。次数は，接続数のより多いノードがグラフの中心となるノードであるという考え方に基づく。有向グラフの場合，次数には，ノードに入ってくる矢印の数である入次数（indegree）と，ノードから出る矢印の数である出次数（outdegree）の2つがある。ノードa_iの入次数を$d_{\mathrm{in}}(a_i)$，出次数を$d_{\mathrm{out}}(a_i)$とするとき，隣接行列Xの要素x_{ij}を用いて

$$d_{\mathrm{in}}(a_i) = \sum_{j=1}^{n} x_{ji},$$

$$d_{\text{out}}(a_i) = \sum_{j=1}^{n} x_{ij}$$

と定義する。

次数ではノード間の接続数のみを考慮するが，接続されるノードの接続状況を再帰的に考慮する指標として，固有ベクトルに基づく中心性（eigenvector centrality）がある。ノードa_iの固有ベクトル中心性v_i（以下，固有ベクトルという）は，

$$v_i = \frac{1}{\lambda} \sum_{j=1}^{n} x_{ij} v_j$$

で定義される。ここで，λは隣接行列の最大固有値である。

2つのノードa_j,a_kに対して，これら2点を最小のステップで結ぶルートを測地線（geodesic）といい，このステップ数を測地線距離という。ノードa_jとa_kを結ぶ測地線の数をg_{jk}とする。別のノードa_iに注目したとき，測地線のうちa_iを通るものの数を$g_{jk}(a_i)$とする。このときノードa_iの媒介性（betweenness）$b(a_i)$を，

$$b(a_i) = \sum_{j<k} \frac{g_{jk}(a_i)}{g_{jk}}$$

と定義する。ネットワークの中に複数のコミュニティ（その中では関係が密であるが，外部とはよりまばらな構造を持つノードの集まり）が存在するとき，媒介性が大きなノードは，コミュニティをブリッジ（媒介）する役割，つまり，コミュニティの領域や機能を拡大する役割を果たしているノードと考えることができる。

2つのノードa_i,a_jの測地線距離を$d(a_i, a_j)$とするとき，ノードa_iの近接性（closeness）$c(a_i)$を，

$$c(a_i) = \frac{n-1}{\sum_{j=1}^{n} d(a_i, a_j)}$$

と定義する。分母はノードa_iから他の全てのノードへの測地線距離の合計なので，近接性が最大のノードは，他のノードへ最も近い距離にあることになる。よって，ネットワークの中心にあると考えることができる。

（2）コミュニティ

いかにしてコミュニティを発見するか。そのために必要なコミュニティ発見のアルゴリズムは，ネットワーク分析の中でホットなテーマの1つである。社会ネットワーク分析の領域でよく知られているものはブロックモデリングである。しかし，Newman and Girvan［8］によるエッジ媒介性を利用する手法が現在，ほぼ標準となっている。また，これを契機として，物理学の分野から様々なアルゴリズムが提案されて来た。その中で，統計物理におけるスピングラスと焼き鈍し法の考え方を適用したものがあり［10］，本稿ではこれらの方法を利用する。

発見したコミュニティ構造の確かさは，モジュール性（modularity）の測度によって評価することができる。ネットワークがk個のコミュニティを持っているとする。e_{ij}を，グループiのノードをグループjのノードに結びつけるエッジの割合とし，大きさ$k \times k$の対称行列を$e = (e_{ij})$と定める。このとき，モジュール性測度Qを

$$Q = \sum_{i=1}^{k} (e_{ii} - b_i^2), \text{ ただし, } b_i = \sum_{j=1}^{k} e_{ij}$$

と定義する。ネットワークの中にコミュニティ構造が全く無いとき，Qは0となる。選択した分割がネットワークの真のコミュニティ構造に近くなればなるほど，Qの値は1に近くなる［8］。

（3）スケールフリー性

ネットワークをマクロに特徴づける性質の1つにスケールフリー性がある。これは，映画の共演ネットワークや細胞代謝ネットワーク，論文の引用ネットワーク等様々な分野で観測されている［1］。スケールフリー性は，例えば次数Xに対して，その確率関数$P(X = x)$を考えるとき，分布がベキ法則に従っている，つまり

$$P(X = x) \sim x^{-\gamma}$$

となることを意味する（γをベキ乗数という）。これの両対数グラフを描くと，γを傾きとして持つ直線となる。スケールフリーという名前は，xの尺度（スケール）を変換して考えても，傾きγには影響を与えないところから名付けられた。ベキ分布の裾の分布$P(X > x)$に関しては，

$$P(X > x) \sim x^{-(\gamma-1)}$$

となるので，確率関数が持つ性質は，裾の分布に対しても成立する。

3．結果

全792本のパッケージの依存関係を調べたところ，339本は，パッケージの依存関係からは独立していた。すなわちR本体のみに依存するものであった。よって，依存関係を持つパッケージは453本となった。453個のノードから構成される依存関係全体のグラフを図3に示す。グラフは，図の中心部に配置されるノード数の多い大きな集団（クラスタ）が1つと，図の周辺部に配置されているノード数が少ない小さな複数のクラスタから構成されている。具体的には，ノード数409のクラスタが1，ノード数4のクラスタが1，ノード数3のクラスタが6，ノード数2のクラスタが11個あった[4)]。

4）既述のように，独立したノード（ノード数1のクラスタ）が339あるが，これは表示していない。

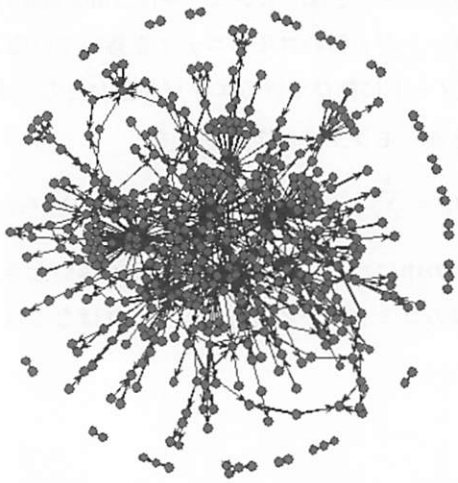

図3　パッケージ全体の依存関係のグラフ

ノード数4以下のクラスタ18個の関係を図4に示す。{maps, mapdata, mapproj} や {wavethresh, CVThresh, Ebayes Thresh} のように，パッケージ名から直感的に関係を把握することのできるグループがいくつかある。他に，{orthopolynom, polynom, kzft} は多項式およびフーリエ解析のパッケージ群であり，{PTaK, tensor} はテンソル解析のパッケージ群であるなど，ほとんどは互いに関連した手法から構成されるクラスタになっている。ノード数3のクラスタに関しては，●→●←●という1点に集中するパターンが5個，●←●→●という1点が他の2点に依存するパターンが1個で，●→●→●や三角形を構成するパターンはない。

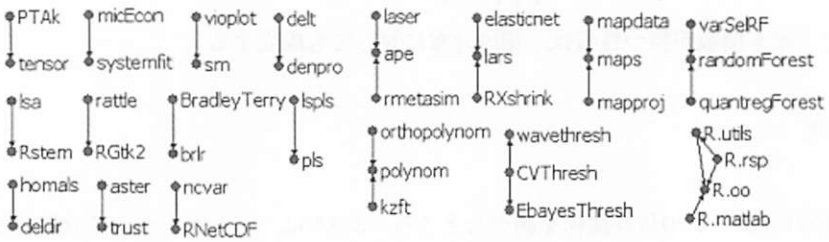

図4　小さなクラスタの依存関係

以下，ノード数409の最大クラスタを分析対象に絞り，ネットワーク指標を用いた詳細な分析を行なう。

入次数および出次数，媒介性のヒストグラムを描くと，図5のようになる。いずれも，右に裾を長く引く，歪んだ形をしている。これは，小さな入・出次数，媒介性のノードが多くあり，大きな入・出次数，媒介性のノードが少数あることを意味する。この特徴は入次数および媒介

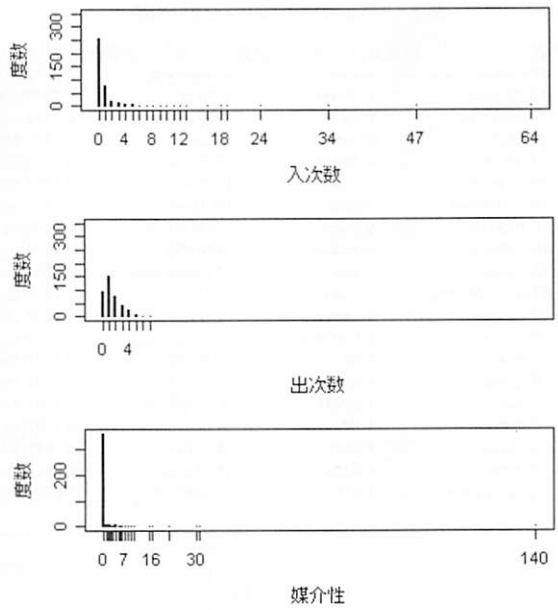

図 5　中心性のヒストグラム：入次数（上），出次数（中），媒介性（下）

性においてより顕著である。入次数の分布関数を両対数軸で図示すると，図 6（左）になる。これより入次数の分布はスケールフリー性を持つことがわかる。また，媒介性の分布もスケールフリー性を持っている（図 6，右）。

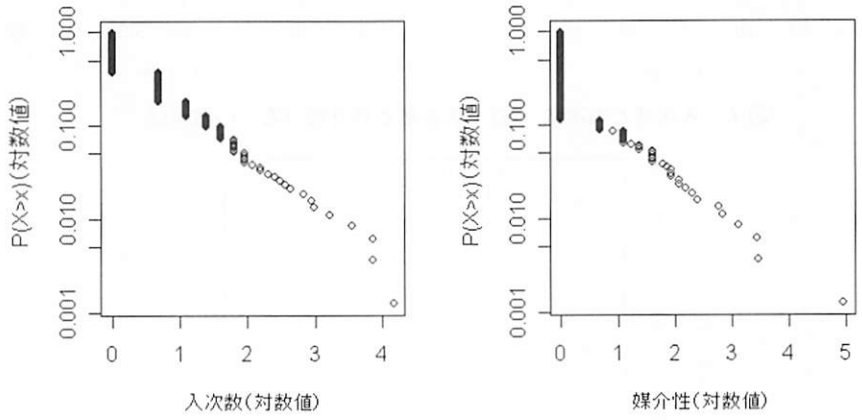

図 6　両対数スケールの分布：入次数（左），媒介性（右）

中心性の指標の大きなパッケージをまとめると表 1 のようになる（上位20）。パッケージの重要性・特徴を見るために，中心性の指標の散布図を作成する。入次数と出次数，入次数と媒介性の散布図を図 7 に，固有ベクトルと媒介性の散布図を図 8 に示す（指標の大きいものにパ

表1 パッケージの中心性指標

No.	パッケージ	入次数	パッケージ	出次数	パッケージ	媒介性	パッケージ	固有ベクトル	パッケージ	近接性
1	methods	64	caMassClass	7	survival	140	methods	0.5552	stats	0.3499
2	MASS	47	fMultivar	6	cluster	31	stats	0.2503	methods	0.3317
3	stats	47	distrTEst	6	boot	30	graphics	0.1449	survival	0.3309
4	survival	34	ipred	6	nlme	21.5	survival	0.1429	MASS	0.3290
5	lattice	24	latentnet	6	Matrix	16	fSeries	0.1402	robustbase	0.3236
6	grid	19	distrSim	5	spdep	15	fCalendar	0.1402	Matrix	0.3129
7	graphics	18	fExtremes	5	coda	10	fBasics	0.1402	graphics	0.3112
8	mvtnorm	16	fPortfolio	5	party	9	fMultivar	0.1356	FLCore	0.3091
9	boot	13	LMGene	5	mclust	8	MASS	0.1324	boot	0.3040
10	nlme	12	pcalg	5	gbm	7	robustbase	0.1252	coin	0.3038
11	tcltk	11	scapeMCMC	5	maptools	7	Matrix	0.1235	lattice	0.3029
12	coda	10	tsfa	5	gamlss	6	distr	0.1202	relaimpo	0.3007
13	cluster	9	survival	4	SemiPar	6	fExtremes	0.1187	pscl	0.2998
14	mgcv	8	Matrix	4	sp	6	fPortfolio	0.1187	SparseM	0.2985
15	utils	8	spdep	4	SparseM	5.5	FLCore	0.1185	sfsmisc	0.2969
16	zoo	7	coin	4	gplots	5	distrTEst	0.1062	portfolio	0.2967
17	class	6	distrEx	4	fMultivar	4	distrSim	0.1062	pcalg	0.2961
18	fBasics	6	FLCore	4	coin	4	fOptions	0.1042	SAGx	0.2959
19	gtools	6	party	4	Zelig	4	sfsmisc	0.0922	its	0.2954
20	sp	6	robustbase	4	e1071	4	portfolio	0.0901	aod	0.2942

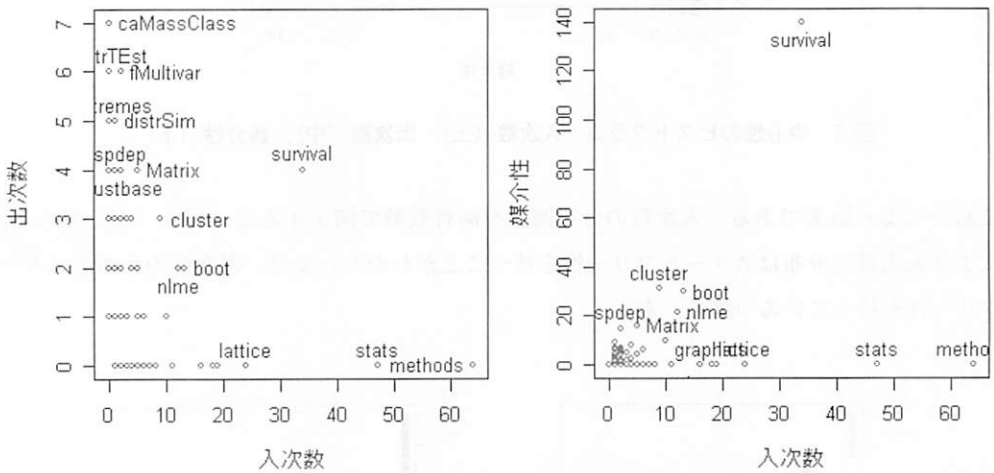

図7 入次数と出次数（左），入次数と媒介性（右）の散布図

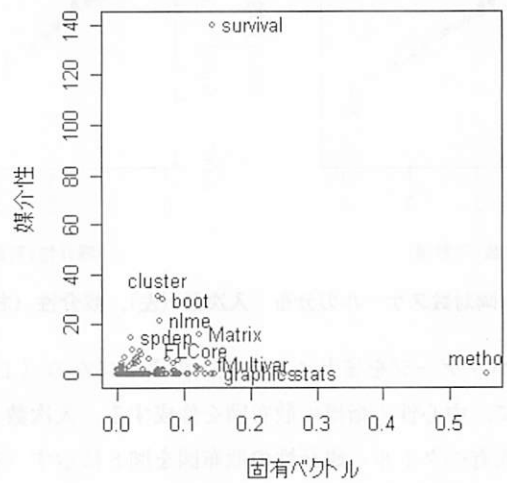

図8 固有ベクトルと媒介性の散布図

ッケージ名のラベルをつけている）。

パッケージには，Rの起動とともにデフォルトで起動される基本的なものがある。それらは，

methods（正式なメソッドとクラス）[5]，stats（Rの統計パッケージ）

graphics（基本となるグラフィックス関数），utils（ユーティリティ関数）

grDevices（Rグラフィックスデバイス，色とフォントのサポート）

の5つであり，基本的にRのコアチームが開発・保守している。また，grid（Gridグラフィックス），tcltk（Tcl/Tkインターフェース）もそうである。grDevicesを除くmethods, stats, grid, graphics, utils, tcltkは，入次数の大きなノードとなっている。さらに，methods, stats, graphicsは固有ベクトルおよび近接性でも大きい。これは，これらがコアチームにより管理されていることからわかるように基本パッケージとして位置づけられるものであり，ネットワークの中で参照されるという意味で中心的な役割を果たしていると解釈できる。よって，以下，これら以外のパッケージについて考察を進める。

図7，8より，MASS（Venables and Ripley［11］の関数とデータセット），lattice（Trelisグラフィックスの実装），mvtnorm（多変量正規分布と多変量t分布の確率，分位点，密度に関する関数）は入次数が大きいノードである（MASSは固有ベクトルでも大きい）。これらに対してsurvival（生存時間解析）は，入・出次数，媒介性，固有ベクトル，近接性のいずれにおいても大きなノードである。boot（ブートストラップ法に関する関数），nlme（線型・非線型混合モデル），sp（空間データのクラスとメソッド）は，入次数，媒介性が大きなノードである。また，Matrix（線形代数に関する関数）は，媒介性，近接性，固有ベクトルで大きなノードである。媒介性が大きなノードとしては他に，spdep（空間統計学），coda（マルコフ連鎖モンテカルロ法の出力の分析と診断），party（再帰分割法）がある。固有ベクトルが大きなノードとしては，fBasics, fCalendar, fSeries, fMultivar, fExtremes, fPortfolio, fOptionsといったRmetrics[6]に含まれる一連の金融工学関連のパッケージが含まれていることが特徴的である。

最大クラスタのグラフを図9に示す。図9では，入次数の大きさに応じてノードの大きさを調整している。この図から明確なグラフ構造を見るのは難しい。そこで，コミュニティ発見の手法を用いて，最大クラスタのグラフにどのようなコミュニティ構造があるかを確認してみる。Newman and Girvan［8］およびReichardt and Bornholdt［10］のアルゴリズムを適用した

5）以下，パッケージ名の後ろの括弧内は，パッケージの説明。

6）Rmetricsは，Swiss Federal Institute of TechnologyのDiethelm Würtzによる金融工学と計算機ファイナンスを教えるためのR環境である。基本的統計学，時間データの管理，時系列データ分析，多変量データ分析，極値データ分析，ポートフォリオ選択・最適化等の関数を含むパッケージ群から成る。

7）このコミュニティ構造の発見はReichardt and Bornholdt［10］のアルゴリズムによるものであり，そのモジュール性測度Qは0.6738であった。Newman and Girvan［8］によると，現実のネットワークに対するQの典型的な値は0.3–0.7の範囲にあるので，この値は大きいと判断できる。

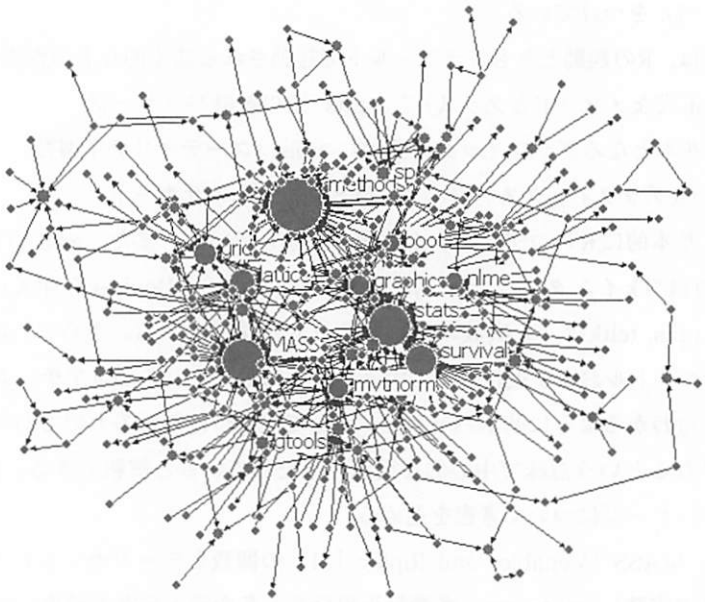

図9　最大クラスタのグラフ：入次数の大きさに応じてノードを拡大

ところ，17コミュニティが発見された[7]。

コミュニティに含まれるパッケージ数（コミュニティの大きさ）を表2に示す。最小のものが大きさ8，最大のものが大きさ44となっている。コミュニティで層別して図に表すと，図10となる。発見されたコミュニティは，図よりうまく分類されていることがわかる。また，意味のある集まりになっている。例えばNo.17（大きさ13）のコミュニティは，パッケージ群｛abind, elliptic, foreign, geoR, gstat, magic, maptools, pixmap, rgdal, rtiff, sp, spgwr, splancs｝より構成されている。geoR, gstat, maptools, rgdal, sp, spgwr, splancsは，空間データ（spatial data）に関するものである。abindは，多次元アレイデータを結合（bind）するための，foreignは，データ分析ソフトウェア（Minitab，S，SAS，SPSS等）のデータやデータベースのデータを読み込むための，rtiffは，tiffフォーマット画像データの読み書きのためのユーティリティである。よって，このコミュニティは，空間データを中心としたデータのハンドリング用ユーティリティ群であると判断できる[8]。

なお，コミュニティNo.1にはnlme，No.2にはgrid, lattice，No.3にはcluster，No.4にはMASS，No.5にはcoda, mgcv，No.6にはfBasics, methods，No.7にはmvtnorm，No.8にはtcltk, utils, No.9にはboot，No.10にはzoo，No.11にはgraphics, stats，No.12にはgtools，No.13にはclass，No.14にはsurvival，No.17にはspという形で，入次数の大きなパッケージがほぼ均等に分散して配置されている。

8）magicは魔法陣作成のユーティリティである。データハンドリング用ユーティリティではないが，abindに依存するため，このコミュニティに分類されている。

表2　コミュニティの大きさ

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
サイズ	23	31	19	13	44	38	26	21	22	13	39	12	22	41	8	24	13

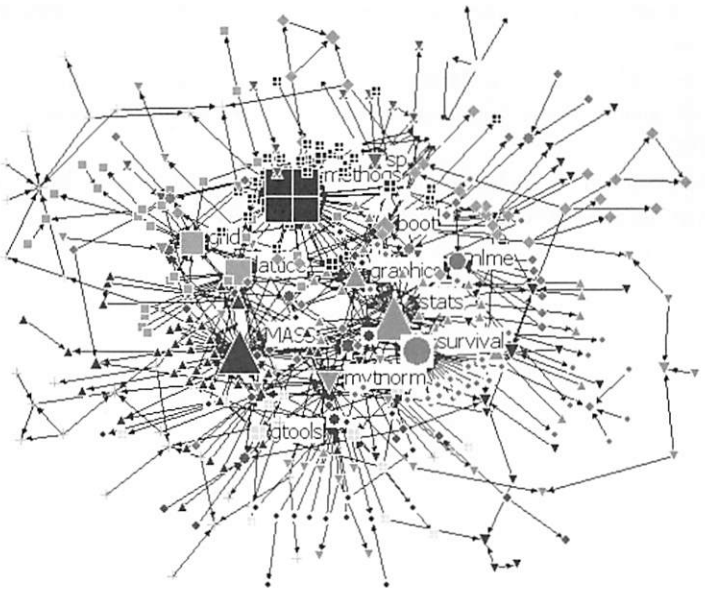

図10　最大クラスタ中の18コミュニティ：ノードの形・色で層別

参考文献

[1] Barabási, A-L. (2002). LINKED: The new science of networks（青木薫訳『新ネットワーク思考—世界のしくみを読み解く』NHK出版, 2002).

[2] Borgatti, S. P. (2002). NetDraw: Graph visualization software, Harvard: Analytic Technologies

[3] Capocci, A., V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli(2006). Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E* 74, 036116.

[4] Crowston, K. and J. Howison (2005). The social structure of free and open source software development, *First Monday* 10, 2, URL http://firstmonday.org/issues/issue10_2/.

[5] Ducheneaut, N. (2005). Socialization in open source software community: A socio-technical analysis, *Computer Supported Cooperative Work* 14, 4, 323–368.

[6] Lopez-Fernandez, L., G. Robles, and J. M. Gonzalez-Barahona (2005). Applying social network analysis to the information in CVS repositories, *1st International workshop on mining software repositories MSR 2004*.

[7] Madey, G., V. Freeh, and R. Tynan (2002). The open source software development phenomenon: An analysis based on social network theory, *Eighth Americas Conference on Information Systems 2002*.

[8] Newman, M. E. J. and M. Girvan (2004). Finding and evaluating community structure in networks, *Physical Review E* 69, 026113.

[9] R Development Core Team (2006). R: A language and environment for statistical computing, R

Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[10] Reichardt, J. and S. Bornholdt (2006). Statistical mechanics of community detection, *Physical Review E* 74, 016110.

[11] Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth edition. Springer.

[12] Wasserman, S. and K. Faust (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

[13] Xu, J., Y. Gao, S. Christley, and G. Madey (2005). A topological analysis of the open source software development community, *Proceedings of the 38th Hawaii International Conference on System Sciences—2005 (HICSS'05)*.

[14] Zlatić, V., M. Božičević, H. Štefančić, and M. Domazet (2006). Wikipedias: Collaborative web-based encyclopedias as complex networks, *Physical Review E* 74, 01611.