日本の特許引用における推移性の検討
～ERGMs（Exponential Random Graph Models）適用の試み～

和田　哲夫

要旨：

1）　経済学・経営学において特許引用はきわめて広範囲に利用されているが、ネットワークとしてみたときのクラスタ性には実証検討が進んでいない。そこで日本特許について、クラスタ性の一側面である推移性（transitivity）を中心的な関心におき、ERGMを含むネットワーク分析を試行した。

2）　日本の審査官特許引用と発明者特許引用それぞれ全体を方向付きネットワークとしてみたとき、後者の推移性指標は、前者よりも高い。

3）　引用ネットワークをマルコフ・グラフとみなし、推移的トリプル数・同一出願人に属する引用ペア数・同一国際特許分類クラスに属する引用ペア数をそれぞれネットワーク指標として扱って、統計分析ソフトウエアR環境上のergmパッケージによる検証を限定サンプルにおいて試みたところ、個々の特許引用に3つの指標は正の影響を与えていると観察された。

1．はじめに

経済学・経営学における研究開発の分析では、特許引用はきわめて広範囲に利用されている（Jaffe and Trajtenberg, 2002）。個々の特許は経済的な情報を十分に含んでいないが、空間的・時間的・組織的に分散した複数の特許間の関係を示す特許引用を通じて情報を集約することにより、経済的側面に関連した分析を行いやすくなるからである。たとえば、企業の特許出願数や特許取得数は、そのままでは企業の利益や時価総額に対してほとんど説明力がなく、せいぜい研究開発に投入される資源の大きさを示す指標として利用価値を持つに過ぎない。しかし、特許引用データを用いて被引用数で保有特許数を加重することにより、企業価値への説明力が上がる、という仮説は複数の実証研究により支持された（Hall, Jaffe, and Trajtenberg, 2005; Nagaoka 2005）。また、特許引用情報は先後関係にある特許間の技術的な関連を示し、後続特許に対する先行特許の影響力を（発明者の主観的認識からみれば必ずとはいえないものの、統計的には）表す、と考えられ（Jaffe, Trajtenberg, and Fogarty, 2000）、知識フローの代理変数としてもしばしば用いられる。

ところで、ネットワークの一種である特許引用にはまだ適用がごく少ないものの、社会学、生物学、コンピュータ科学など各種分野にまたがる「ネットワークの科学」が物理学を中心として確立しつつある。ネットワークを構成する各頂点間の平均距離が小さくなるスモールワールド性（Watts and Strogatz, 1998）や、各頂点の接続数分布がベキ則に従うスケールフリー性（Barabasi

and Albert, 1999）などが、文献引用ネットワークを含む多くのネットワークに観察されることが知られるようになった。このうちスモールワールド・ネットワークに付随する重要な性質として、１つの頂点を共有する２つの辺は、別の１辺によって結ばれやすいという意味でのクラスタ性があり、人間社会に関係したネットワークには広く観察されることも知られている。

このようなネットワーク科学の隆盛にも関わらず、経済学・経営学に特許引用データを応用する上で、クラスタ性などネットワークの諸指標に対して現在までは十分な注意が払われていない。たとえば、中小企業の自社内引用（自己引用：self-citationともいう）比率の上昇は企業価値にプラス、大企業では影響なし、またはマイナスの影響、という観察を行った研究がある（Hall, Jaffe, and Trajtenberg, 2005）。この前提として、企業の特許数と自己引用数が比例的に増加する想定がある。つまり引用ネットワークが基本的にはランダム・ネットワークであるという想定に立っている。しかし引用ネットワークにおいてローカルクラスタを形成する技術的・人的な理由がある場合、企業内特許が一定数を超えるまではローカルクラスタが一企業内には生成しにくい、というような現象が存在するかもしれない。そのような現象があれば、ランダム・ネットワークの仮定は特許引用の経済価値を議論する前提として歪んでいることになってしまう。クラスタ性が特許引用ネットワークに存在する蓋然性は、他の文献引用ネットワークの研究から高いものと思われるが、引用のクラスタ性の程度について把握した研究が未だほとんどみられない。

このような動機に基づき、特許引用ネットワークのクラスタ性について初歩的な分析を行う。ここでは、あまり検討がなされていない日本特許について、推移性（transitivity）を中心的な関心としつつERGMを用いたネットワーク分析までを試行した。

以下では、まず第２節で、日本の特許全体について特許引用ネットワークの推移性指標を計測し、審査官引用と発明者引用の間の差異を示す。第３節では、近年発展してきたERGM手法の概要を示し、クラスタ性が企業内外・技術分野だけでは説明できない普遍的な性質の一つなのかどうかを考えるため、ごく限られたサンプルながら実際のデータを用いた分析を試行した結果を示す。第４節では、試行した結果を踏まえて意義と課題について議論する。

２．推移性指標

２－１．推移性指標

まず、引用ネットワーク全体の推移性指標を「ある特許から他の特許へ２つの引用関係を通って到達する経路、すなわち長さ２の直列的な引用経路を構成する３つの特許のうち、端点にあたる２つの特許の間に直接の引用関係も存在するものの割合」と定義する。ある特許Xを引用する特許Aと、Xから引用される特許Bとの間が、AがBを直接引用する別の引用関係によって結ばれている、つまり推移的トリプルを形成している、ということなので、上に述べたクラスタ性を表す一つの指標となる。なお、特許の引用関係では、何らかの制度上の特殊要因がない限り古い出願番号を持つ特許を若い出願番号を持つ特許が引用する。上記の特許Aと特許Bとの間が、BをAが引用する関係によって結ばれているような循環的（cyclical）な引用は、通常は起こらないケースなので、このあとの実際のデータ処理において除外(1)して考えている。

(1)　このあとのデータで実際に観察されるトリプルのほとんどが推移的トリプルであるが、引用関係の全体の約

2－2．日本特許引用全体データの推移性

ここでは、特許庁による整理標準化データをもとにした審査官特許引用データと、特許公報に掲載されている引用を集めた発明者引用の２種類を対象とした(2)。審査官引用と、発明者特許引用全体の和集合をとったとき、全体で7,676,949特許からなる引用関係が15,790,114ペア存在している。うち約55.3%にあたる8,736,360引用ペアが審査官引用、約52.3%にあたる8,260,123引用ペアが発明者引用で、全体では審査官引用と発明者引用はほぼ半数ずつと観察される。重なって審査官引用かつ発明者引用となる引用関係が含まれているが、全体の約7.6%と、あまり大きな割合ではない。

統計解析ソフトのＲ上で動く社会ネットワーク分析パッケージsnaでは、推移性を関数gtrans()によって求めることができる。これを日本特許引用データ全体に適用すると0.08511213と算出された。すなわち長さ２の直列的な引用経路を構成する３つの特許のうち、端点にあたる２つの特許の間に直接の引用関係も存在するものの割合は約8.5%となる。これを審査官引用だけとってみれば0.05462885、発明者引用だけとってみると0.1206631となる。審査官引用には約5.5%に推移的なトリプルが存在するのに対して、発明者引用では約12.0%と、２倍以上も多い割合でトリプルが存在することがわかる。

審査官引用と発明者引用の違いについて、米国の特許引用では多くの研究がなされている。たとえば、引用関係にある２つの特許間の発明者住所の距離分布についていえば、審査官引用と発明者引用はあまり異ならない、という結果を報告した研究がある（Alcacer and Gittelman, 2006）(3)。しかし米国特許の引用においても、日本特許にみられるようにクラスタ性に大きな差異があっても不思議ではない。また、当該研究では、企業の枠を超えた引用が発明者ではなく審査官によってなされる傾向があることも報告されているが、発明者はクラスタ性の高い引用を行う傾向があるためではないか、という可能性には触れられていない。このほか、審査官による引用数の方が発明者による引用よりも特許価値の説明力が高い、という報告もある（Hegde and Sampat, 2009）。その原因については議論がなされていないが、ローカルクラスタを超えた人的関係は経済的な価値が高い、という議論は古くから存在し（Granovetter, 1973; Rauch, 2010）、低いクラスタ性を持つ審査官引用が高い経済価値との相関を持つという先行研究との類似性を示唆する。

2－3．日本特許の90年代サンプル中心とした特許引用データの推移性

ここまでは、日本特許引用の全体についての平均的な推移性をみたが、1990年代以降の高い特許出願数に比べ、比較的には出願数の少なかった1970年代以前を含むデータは、異なったネットワーク構造を有している可能性を否定できない。そこで、経済産業研究所により2007年に行われ

0.3%に逆の番号関係がみられ、特殊要因の影響を除くため分析サンプルから除外している。

(2) 審査官引用は鈴木潤教授による2007年までの整理標準化データ、発明者引用は玉田俊平太教授・内藤祐介氏による2008年までの特許公報ベースのデータを利用。

(3) この研究では、転職した技術者に特有の引用インセンティブが訴訟危険のため働くであろう、など他の点が議論されている。

た発明者サーベイ（長岡・塚田、2007）で分析対象となっている5,278特許[4]を中心として、前後3世代[5]までの引用関係（審査官または発明者引用）にある特許の和集合を母集団にとってみる。この中では1,185,399の特許が特定でき、これらは3,260,563ペアの引用関係を形成している（全体の約20.6%に相当）。ペアワイズに数えたとき、この発明者サーベイ特許から距離3以内の引用関係のうち73.18%を発明者引用が占め、審査官引用は31.7%である。引用距離の意味において近距離にある範囲に、発明者引用は審査官引用の倍以上の数が存在することがわかる。

上記の発明者サーベイ特許から距離3以内の引用ネットワークでは、推移性指標は 0.0984であった。審査官引用関係に限定したとき、推移性指標は 0.0390であり、発明者引用関係に限定すると0.1166である。つまり、発明者引用はやはり12%近くの推移性を持ち、4%に満たない審査官引用よりもはるかに推移性が高いことがわかる。密度（ネットワークにおいて張ることができるすべての辺に対する実際の辺の比率）も、全体では2.30×10^{-6}、審査官引用では7.52×10^{-7}、発明者引用は1.69×10^{-6}なので、やはり発明者引用の方が倍以上高い数値となっている。

いずれの指標からも、発明者引用のほうが審査官引用よりもネットワーク的にみてローカルに密集していることがわかる。引用関係がスモールワールド性を持つとすれば、個々の引用関係の存在確率は、ネットワーク頂点としての特許を共有する他の引用関係の存在と統計的に相関している可能性が高い。引用の中でも、発明者引用ではクラスタ形成確率が高い、という意味で相互依存の影響を受けやすいと思われる。クラスタは同一企業内に属する研究者間で相互に特許引用が付されやすい、また同一技術分野内で相互引用が付されやすい、という条件によっても形成されると思われるが、このような企業内・技術分野内の集合的性格によってのみクラスタは形成されるのだろうか。次節ではこの点について検討する。

3．ERGMs

3－1．モデル概要

統計物理を応用したネットワーク統計的モデルは、次数分布に関するベキ法則などの少数の基本的性質に関心を集約し、この10年余りの間に急速な発達を見せてきた。しかし実際のネットワークの構造に対して、モデルを推定し、現実のデータに対する適合を実際に検定する統計的な分析手段はまだそれほど多くない（Goldenberg et al, 2009）。とくに、推移性のモデル化や検定は厄介である（Snijder et al. 2006）。その中で、マルコフ・グラフとしてのモデル化手法（Frank and Strauss, 1986）を一般化したERGMs（Exponential-Family Random Graph Models）またはp^*と呼ばれるモデル（Wasserman and Pattison, 1996）は、推移性を含めた様々な性質のモデル化と検定への道を開いた。この文法では、ネットワーク全体をマルコフ・グラフと仮定して、注意を向ける個々の辺の発生確率を、それを含まないネットワーク全体の構造によるexponentialな関数として記述する。この関数には最尤法が適用可能なため、統計解析ソフトのR上で動くstatnet（ergmを含む）や、SIENAというパッケージを通じて実際にデータに基づいて検定をすることが可能な

(4) 主に90年代後半の日米欧3極出願であり、特許全体からすると重要な部分集合からサンプルをとって発明者に直接アンケート回答を求め、分析可能な回収内容となった集団を指している。

(5) 単純に引用を3世代さかのぼった特許や、3世代まで被引用を下った特許のほか、引用パス長が3以内（geodesic distance≦3）の特許すべてをとった。

状態に達している（Hunter et al, 2008; Morris et al, 2008）。

ここでは、上述した発明者サーベイ回答に含まれる電子部品製造業２社の12個の特許に着目し、そこから引用距離３以内の特許すべて（ただし引用側が被引用側よりも古い番号の引用ペアを除く）の和集合からなるネットワークを試行的に分析対象とする。上記のergm分析パッケージは、数百万以上の特許引用ネットワークのような巨大なネットワークデータには計算機資源上の制限が理由でまだ適用不可能なので、このような少数サンプルを対象に限定した。ここでは、実際には777ペアの引用関係が分析対象となる。この引用関係は546個の特許から構成され、そのうち62.7%が審査官引用、40.9%が発明者引用である。もともと12個の異なる特許を中心としてサンプリングしているため、引用はすべてが互いにつながりをもっているわけではないが、図示するとかなりの局所的な集積が図１からみてとれる。

３－２．推定方法

R上で動くergmパッケージでは、引用ネットワークデータに対して①引用関係について出願人が同一である辺の数、②引用関係について国際特許分類のクラスが同一である辺の数、③ネットワークに存在する推移的トリプルの数、の３つの影響を検討するために、次のようなコマンドを用いることができる。

```
fullmodel <- ergm (n7 ~ edges + nodematch ("applicant1", diff = F)
+ nodematch ("ipc4id", diff = F) + ttriple)
```

ここで、fullmodelは推定結果を格納するオブジェクトの名称、n７は使用したサンプルデータからなるネットワークオブジェクトの名称で、それぞれアドホックなものに過ぎない。"ergm"がERGMモデル命令を表し、"edges"、"nodematch"、"ttriple"が適用するネットワーク統計量を示す。"edges"はネットワーク内の辺の数を意味する。"nodematch"は、diff変数を"diff=FALSE"のように指定したとき、括弧内で最初に指定されたノード属性が同一の辺のカウント量を意味する。ここでは、第一出願者コード（"applicant１"）と国際特許分類の最初の４桁で表される「クラス」（"ipc４id"）のそれぞれが同一である辺の数を変数として加えており、同一のグループに属する数が説明力を持つかどうかをみようとしている。最後に、"ttriple"が推移的トリプル（"transitive triple"）のネットワーク内の存在量を示す。

３－３．実行結果とその解釈

実行結果は次のようになり、当然に説明力を持つネットワーク内の辺の数のほか、出願者の同一性、特許分類（「IPCクラス」）の同一性や、推移的トリプルが統計的に有意な説明力を持っているように解釈できる。

Monte Carlo MLE Results:

	Estimate	Std. Error	MCMC s.e.	p-value
edges	-6.89169	0.06278	0.005	< 1e-04 ***
nodematch.applicant 1	0.61850	0.18465	0.005	0.00081 ***
nodematch.ipc 4 id	2.75117	0.08412	0.007	< 1e-04 ***
ttriple	0.70141	0.10261	0.007	< 1e-04 ***

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05

ここでは限られたサンプルに基づく単なる試行結果であることに注意しなければならないが、いくつかの変数の組み合わせやサブサンプルから、ここで示した他に次のような結果が得られている。

まず、ここで関心の中心となっている出願者の同一性、特許分類（IPCクラス）の同一性や、推移的トリプルの３つのネットワーク指標は、それぞれを"edges"変数とだけ組み合わせて説明変数として用いたとき、審査官引用のみ、発明者引用のみ、またはその双方のすべてのサンプルの組み合わせにおいて常に有意性を示す。３つのネットワーク指標のうち２つ以上を組み合わせ、サンプルを様々に変化させたとき、特許分類IPCの同一性と推移的トリプルは常に有意のままだが、出願人が同一である数は有意でなくなる結果も一部に観察された。

ここで一貫してみられる結果は、推移的トリプルの頑健な有意性である。つまり推移的トリプルがネットワーク構造として個々の引用生成確率に正の影響を与えている、と解釈可能である。また、特許分類クラスによって定義された技術分類の同一性も、引用確率に対して頑健に正の影響を与えているようにみえる。一方、出願者が同一であることは、変数の組み合わせによっては有意性がみられなくなる場合がある。この場合にも、推移的トリプルは一貫して有意性を失わないので、企業内でクラスタが形成される傾向はあるが、より一般的なクラスタ形成原理に支配されている可能性がある。

３－４．分析上の問題点

以上は、サンプル数が限定されている中での試行結果であり、他にも手法上、またデータ上の様々な制約が存在している。分析手法についての最大の問題として、ERGMの単純な適用により、ネットワークを静的な系として扱っていることが挙げられる。特許引用は、本来は時系列に沿って成長するネットワークであり、個々の引用付加に際しては、その引用時点以前に存在したネットワーク構造に規定される、という前提に立つべきだが、ここではそのような動的モデル化ができていない。言い換えると、古い引用の付加確率は、その後に付加された若い引用のネットワーク構造にも規定されている、という基本的な考え方に立っているという問題がある。また、ergmパッケージでは、推定にあたってマルコフ連鎖モンテカルロ法を用いるため、結果は推定するごとに若干異なりうる、という問題もある。

この他、個人IDが利用可能でないため、同一発明者が関与した特許かどうか、また同一審査官が関与した特許引用かどうか、というような変数が利用できていない。出願人情報は、整理標準

化データにもともと含まれている出願人コードを直接利用しているが、この場合は企業の名寄せが不十分となる欠点があり、また筆頭出願人しか今回は用いていないので、共同出願は無視してしまっている。特許分類情報も、同じIPCクラスに属するかどうかのバイナリ情報しか使っておらず、より細かいグループやサブグループ情報を利用していない。また一特許にふつうは複数の分類が付されているという現実を捨象してしまっている。さらに、グラフの方向として、先行特許から後続特許への影響という時系列側面を重視するために引用側をhead、被引用側をtailと入力しているが、引用する行為に着目した方向性とは逆なので、これがどのような構造の違いを生んでいるのか、現時点ではわかっていない。このような様々な課題は残されている。

4．今後の可能性および課題

ここでの検討は、引用ネットワーク全体の大まかな把握と、ERGM手法の試行までにとどまる。このような範囲でも、先行研究の提起した具体的な問題の一部にも手がかりを与えている。その一つとして、審査官引用と発明者引用の間でネットワーク構造が根本的に異なる、という事実が挙げられる。米国における審査官引用と発明者引用は、住所距離の点からは差異が少ない、というAlcacerらの研究に関しては、引用の由来が基本的に異なるのではないか、という当然の事実を、今までにはあまり利用されていなかったネットワーク構造上の特徴とともに指摘できる。また、経済産業研究所の発明者サーベイにおいて、審査官引用で測定しても、企業内からの引用数のみが先行特許への依拠肯定の確率と相関を持ち、企業外からの引用数は説明力がない、という問題(6)が明らかになっていたが、審査官引用にもクラスタ性が広汎に観察される、という今回得られた事実が説明理由として考えられる。

以上のように、従来あまり用いられていない統計分析手法の登場により、特許引用の意味について新たな視点と分析可能性が生まれたといえる。技術距離、引用ラグ、地理的距離などが引用確率に影響を与えることは既知だが、企業組織内外との関係や、その他のネットワーク構造との関係は十分解明されていないので、大きな課題群が登場したということもできる。これと関連して、オリジナリティ指標などにも今後の進歩が期待できる。ただ、処理可能なサンプル数を増やす計算機論上の困難性や、動的なネットワーク分析手法が未発達であることなど、簡単ではない問題もいくつか残されており、分析手法そのものの進歩も必要とされていることは確かである。

参考文献

Alcácer, Juan, and Michelle Gittelman, 2006, Patent Citations as a Measure of Knowledge

(6) 発明者サーベイの一次サーベイにおいては、先行技術への依拠や、先行特許の存在について特許発明者からの回答を得た。企業内・外からの特許引用数と、「先行特許に基礎をおいていた」と回答する確率の関係を探った結果、先行特許がある、と回答した発明者について、企業組織の内外からの特許引用数と、発明者が認識する先行特許の社内外の区別認識は良く一致する。また、企業内からの特許引用数が多いとき、発明者の先行技術が存在した、という認識ともよく一致した。しかし企業外からの特許後方引用数は、発明者の認識とは統計的な関係がみられなかった。結果として、企業内外をあわせた後方引用特許の総数は、「先行特許に基礎をおいていた」という回答確率との関係も見いだせなかった。発明者引用での結果ならばともかく、審査官引用で測定しても、企業内引用数のみが先行特許への依拠肯定の確率と相関を持つ、という結果が得られることに対して説得的な理由は、従来の枠組みからは考えがたい（和田、2010）。

Flows: The Influence of Examiner Citations, *Review of Economics and Statistics*, Vol. 88, No. 4, pp. 774-779.

Barabási, Albert-László, and Réka Albert, 1999, Emergence of Scaling in Random Networks, *Science*, vol. 286. no. 5439, pp. 509-512.

Frank, Ove, and David Strauss, 1986, Markov graphs, *Journal of American Statistical Association*, Vol. 81, no. 395, pp. 832-842.

Granovetter, Mark S. 1973, The Strength of Weak Ties, *American Journal of Sociology*, vol.78, no.6, pp.1360-1380.

Hall, Bronwyn, A. Jaffe, and M. Trajtenberg, 2005, Market Value and Patent Citations, RAND *Journal of Economics*, vol.36, no.1, pp.16-38.

Hegde, Deepak, and Bhaven Sampat, 2009, Examiner citations, applicant citations, and the private value of patents, *Economics Letters*, vol.105, no.3 pp.287-289.

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris, 2008, ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks, *Journal of Statistical Software*, Vol.24, nol.3.

Jaffe, Adam B., and Manuel Trajtenberg, 2002, Patents, Citations, and Innovations: A Window on the Knowledge Economy, MIT Press.

Jaffe, Adam B., Manuel Trajtenberg, and Michael S. Fogarty, 2000, Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors, *American Economic Review*, vol. 90, no. 2, pp.215-218.

Goldenberg, Anna, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi, 2009, "A Survey of Statistical Network Models," in *Foundations and Trends in Machine Learning*, Vol.2, No.2,

Morris, Martina, Mark S. Handcock, and David R. Hunter, 2008, Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects, *Journal of Statistical Software*, Vol.24, nol.4.

Nagaoka, Sadao, "Patent Quality, Cumulative Innovation and Market Value: Evidence from Japanese Firm Level Panel Data," 2005, *IIR Working Paper* WP#05-06.

Rauch, James E., 2010, Does Network Theory Connect to the Rest of Us? A Review of Matthew O. Jackson's Social and Economic Networks, *Journal of Economic Literature*, Vol.48, no. 4, pp.980-986.

Snijders, Tom A.B., Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock, 2006, New specifications for exponential random graph models, *Sociological Methodology*, Vol. 36, no. 1, pp. 99–153.

Wasserman, Stanley and Philippa Pattison, 1996, Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p, *Psychometrika*, Vol.61, no.3, pp.401-425.

Watts, Duncan J. and Steven H. Strogatz, "Collective dynamics of 'small-world' networks," 1998, *Nature*, vol.393, pp.440-442.

長岡貞男・塚田尚稔, 2007,『発明者から見た日本のイノベーション過程：RIETI 発明者サーベイの結果概要』RIETI Discussion Paper Series 07-J-046.

和田哲夫, 2010,『発明者による先行特許認識と特許後方引用』, RIETI Discussion Paper Series 10-J-001.

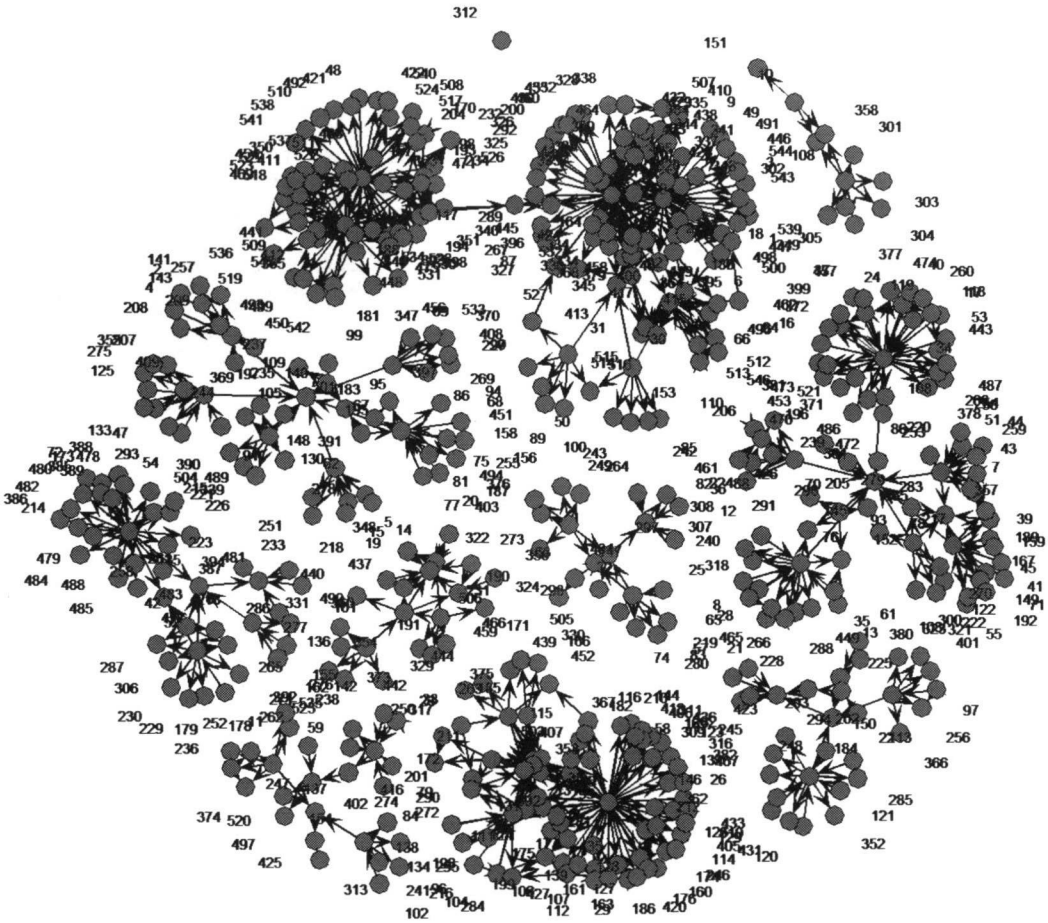

図 1