doi:10.11959/j.issn.1000-436x.2015025

复杂网络中 *k*-核与网络聚集系数的关联性研究

刘君，乔建忠
（东北大学 信息科学与工程学院，辽宁 沈阳 110819）

摘 要：选取复杂网络特征变量—聚集系数为研究目标，通过数学推导与证明，清晰描述了 *k*-核与聚集系数的关联性。通过仿真实验证明，随着 *k*-核的不断解析、*k* 值的不断增加，网络聚集系数亦呈现逐步增加的趋势。该结论为 *k*-核解析在复杂网络中的进一步应用提供相应的理论基础与指导。
关键词：复杂网络；*k*-核；聚集系数；关联性
中图分类号：TP393 **文献标识码**：A

Research on relevance between *k*-core and clustering coefficient in complex network

LIU Jun, QIAO Jian-zhong
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China)

Abstract: *K*-core analysis is an effective way to simplify the graphic topological structure. Many researches considered that the higher value *k* is, the more important the core is in complex network. But the relevance analysis between *k*-core and clustering coefficient has not been made. Experimental results show that with the *k*-core analysis, the trend of the clustering coefficient is consistent with *k*. The proposed conclusions can provide theoretical basis and guidance for the future applications of *k*-core analysis in complex network.
Key words: complex network; *k*-core; clustering coefficient; relevance

1 引言

复杂网络是一种新兴的网络研究理论，该理论正渗透到数理学科、生命学科和工程学科等众多不同的领域。学术界关于复杂网络的研究方兴未艾。特别是，国际上有 2 项开创性工作掀起了一股不小的研究复杂网络的热潮。一是 1998 年 Watts 和 Strogatz 在 Nature 杂志上发表文章，引入了小世界(small-world)网络模型，以描述从完全规则网络到完全随机网络的转变；二是 1999 年 Barabási 和 Albert 在 Science 上发表文章指出，许多实际的复杂网络连接度分布具有幂律形式[1,2]。近些年，对于复杂网络的研究出现了较多成果，大致可以分为 2 类：1）结构分解类：理解真实世界中复杂网络（如 Internet，WWW，细胞组织网络等）的体系结构并抽取出它们中紧密联系的部分——社团、结构洞、*k*-核等，发现它们之间的联系[3,4]。2）特征指标类：研究人员提出了一系列复杂网络特征（如介数、度分布、聚集系数等）来描述真实世界中复杂网络的拓扑结构。这些指标为宏观统计学角度研究复杂网络提供了非常有力的工具，例如通过统计 Internet 中节点的数据交互，分析获取 Internet 拓扑结构的介数、度分布、聚集系数等特征，依据这些统计特征就可以设计相应的实模来仿真 Internet 的拓扑结构。

总体来看，目前关于复杂网络的研究多数只是停留在宏观统计分析上，通过对某一事件的关联数据进行统计，采用曲线估计、拟合等方法分析并找出规律，例如在文献[5]中，作者就 Internet 的拓扑结构突发性改变展开研究，通过统计大量 Internet 的数据来尝试解释突变的原因。宏观统计

收稿日期：2013-08-10；**修回日期**：2013-12-05
基金项目：国家自然科学基金资助项目(61273071)
Foundation Item: The National Natural Science Foundation of China (61273071)

分析能够从统计学角度解释很多一直困扰让人的问题，但是由于数据的偶然性和时间的局限性，很难确保统计的样本量对于规律描述是充分的，此时需要一种科学客观的复杂网络拓扑结构描述理论。将宏观统计分析与该结构描述理论相结合能够更科学地解释一些现象，然而目前对于复杂网络拓扑结构描述理论方面的研究几乎处于空白。*k*-核解析是一种高效的图形分析方法，通过该方法，网络逐渐趋于核心的区域，一些研究人员给出了定性的结论：越中心的核，连通性越强。但是为什么会出现这种规律，连通性增强的幅度等问题均未回答。为此本文从 *k*-核解析模型着手，研究了其数学意义，推导分析了该模型与网络聚集系数的关联性。得出的相关结论能够为 *k*-核解析的进一步应用奠定相应的基础。

2　*k*-核解析模型

设图 $G=(V, E)$是由$|V|=n$个节点和$|E|=e$条边所组成的一个无向图，则 *k*-核的定义[6]如下。

定义 1　*k*-核（*k*-core）。由集合 $C \subseteq V$ 推导出的子图 $H=(C, E|C)$，当且仅当对 C 中的任意节点 v，其度值均大于或等于 k，即 $\forall v \in C$: degree $H(v) \geqslant k$，具有这一性质的最大子图就叫作 *k*-核。核中包含的节点数目则称为核的大小。依据 *k*-核的定义，图 G 的 *k*-核就可以通过反复地移去那些度值小于 k 的节点以及与其连接的边，直到余下图中所有节点的度值都大于或等于 k 来得到。因此可以通过 *k*-核解析由外层至内层一层一层地解析网络，直到最内层为止，从而揭示网络的层次结构性质。

图 1 所示为 *k*-核解析的流程。首先按照 *k*-核定义去掉图 1 中度值为 1 的黑色节点及其边，剩下的由灰色与白色节点构成的拓扑图即为 2-核；然后去掉度值为 2 的灰色节点，剩下的由白色节点构成的拓扑结构图为 3-核。需要注意的是，若图 1 中出现 Y_4 节点，按照核解析的定义，需要反复去掉度值为 2 的节点，因此$\{Y_1, Y_2, Y_3, Y_4\}$都将会在 2-核解析过程中被去除，即虽然它们初始度值不同，但是最终都属于 2-层节点。

3　*k*-核与聚集系数关联性分析

大量文献给出了关于 *k*-核解析的定性描述：高核内节点的连通性高、传播性强。但是连通性、传播性到底代表什么，用什么来表示，均未给出详细定量的描述。本文将连通性、传播性统称为小世界特性，并引入网络直径与聚集系数概念来表征网络小世界特性的强弱。网络直径越小、聚集系数越大的网络小世界特性越强，反之越弱。

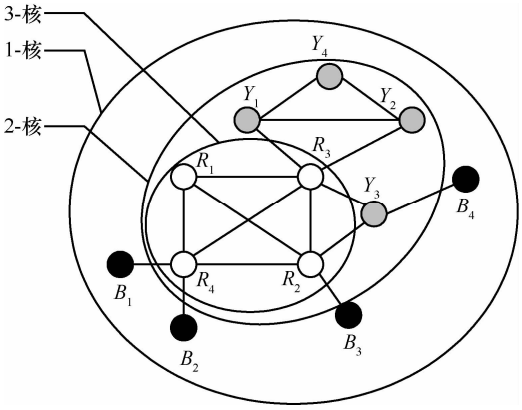


图 1　*k*-核解析示意

定义 2　节点聚集系数(clustering coefficient)[7~9]。某节点 i 的聚集系数为该节点所有邻居节点之间连接数目占可能的最大连接数目的比例

$$CC_i = \frac{2E_i}{l_i(l_i-1)} \tag{1}$$

其中，l_i 为节点 i 的邻居节点个数，E_i 为这 l_i 个节点之间存在的连接数目。一个网络的聚集系数为网络中所有节点 CC_i 的均值 C。节点聚集系数是反映节点在网络连通性特征上贡献的一个重要指标。网络聚集系数是反映网络拓扑结构连通性、传播性的一个关键因素。

定义 3　网络直径。指网络中任意两节点间跳数的最大值[10,11]。网络直径与聚集系数共同确定着网络小世界特性的强弱。例如图 2(a)网络的网络直径为 6 跳，网络聚集系数 C_a 为

$$C_a = \frac{CC_{R_1}+CC_{R_2}+CC_{R_3}+CC_{R_4}+CC_{Y_1}+CC_{Y_2}+CC_{Y_3}}{7}$$

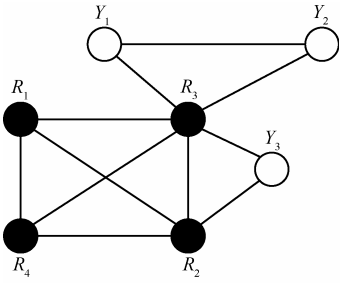
$$= \frac{1+\frac{2}{3}+\frac{1}{3}+1+1+1+1}{7} = \frac{6}{7} \tag{2}$$

图 2(b)中网络直径为 3 跳，网络聚集系数 C_b 为

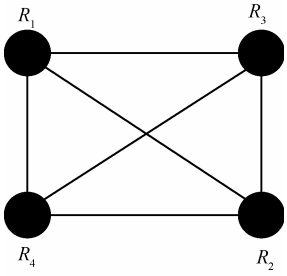
$$C_a = \frac{CC_{R_1}+CC_{R_2}+CC_{R_3}+CC_{R_4}}{4} = \frac{1+1+1+1}{4} = 1 \tag{3}$$

由此可以看出图 2(b)网络的小世界特性较图 2(a)网络的小世界特性强，即连通性、传播性均要

高于图 2(b)网络。因此在图 1 给出的网络拓扑基础上进行 *k*-核解析满足上文提及的“高核对应着高连通性与传播性”结论。但是是否在任意给定的拓扑结构中这一结论均成立？下文将围绕这个问题对命题 1 展开论证。



(a) 网络直径为 6



(b) 网络直径为 3

图 2 *k*-核解析局部示意

命题 1 在给定网络拓扑上，不断进行 *k*-核解析，若 *k* 核对应的网络聚集系数为 C_k，*k*+1 核对应的网络聚集系数为 C_{k+1}，则有 $C_{k+1}>C_k$；若 *k* 核对应的网络直径为 D_k，*k*+1 核对应的网络直径为 D_{k+1}，则有 $D_k<D_{k+1}$。

证明 首先关于网络直径的变化：由 *k*-核解析的定义可知，随着核数的增加，*k* 核变为 *k*+1 核时，网络中节点数目是减少的，且节点间的连边也是只有减少没有新增。所以网络中两点间最大跳数不可能增加，即 $D_k<D_{k+1}$。其次关于聚集系数的变化，对式(1)进行变形

$$CC_{i(k)}=f(E_{i(k)})=\frac{2}{l_i(l_i-1)}E_{i(k)}=K_{i(k)}E_{i(k)} \quad (4)$$

在任意结构中，*k*-核内节点 *i* 的邻居节点数 k_i 为常数，能影响 *k*-核结构中节点 *i* 聚集系数的因素为 $E_{i(k)}$。此外通过式(4)可以看出 *k* 核中节点 *i* 的聚集系数 $CC_{i(k)}$与 *k* 核拓扑中节点 *i* 的邻居节点间连边数目呈离散线性方程关系，其中，$K_{i(k)}$为 *k* 核中节点 *i* 对应方程系数。

图 3 为拓扑由 *k*-核向 *k*+1-核解析的示意，*k*-核内节点 *I*、*J* 的度值为 $d_{J(k)}$与 $d_{I(k)}$，且 $d_{J(k)}<d_{I(k)}$。节点 *I* 为 *J* 的邻居节点，在该次解析过程中按照 *k*-核解析的原则将被去除，且 *J* 节点能够保留在更高核内。当节点 *I* 被去除时，对于拓扑中其他剩余节点对应聚集系数将会有 2 种结果：保持不变或改变。

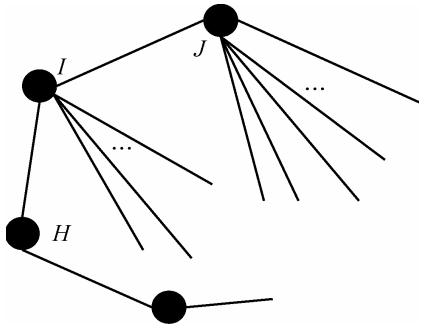


图 3 *k*-核向 *k*+1-核解析示意

本文主要讨论节点 *I* 被去除后聚集系数发生变化的节点。例如以图 3 中的节点 *J* 为例，节点 *I* 为 *J* 的邻居节点，且节点 *I* 与 *J* 的其他邻居节点存在连边，则 *I* 的去除将会影响节点 *J* 的聚集系数。由于给定拓扑的聚集系数与拓扑内部连边数目呈过原点的线性关系，如图 4 所示，因此，只需要确定式(4)线性方程中的 $K_{i(k)}$即可确定 *k*-核与 *k*+1-核对应的线性图。从图 4 中可以看出，若 *k*-核解析后节点 *J* 的邻居节点间连边数目相等，即 $E_{J(k)}=E_{J(k+1)}$，则 *k*+1 核中节点 *J* 对应的网络聚集系数大于 *k* 核中节点 *J* 对应的网络系数。

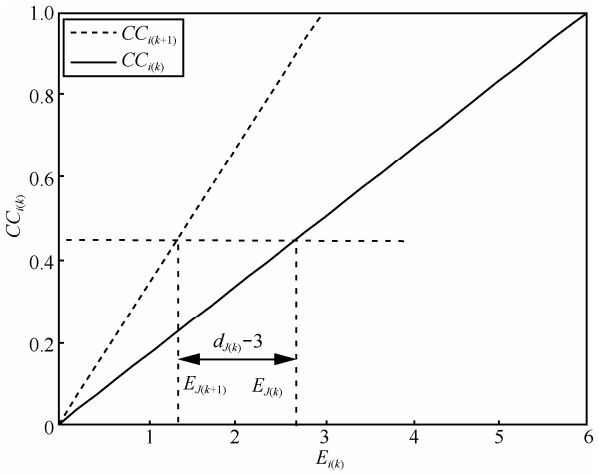


图 4 聚集系数与 E_i 的线性方程

然而由于 *k*-核解析，*I* 节点被去除后，导致了 *k*+1 核中连边数目发生了减少，即 *I* 被去除后，$K_{J(k)}<K_{J(k+1)}$，但是 $E_{J(k)}>E_{J(k+1)}$，难以确定最终节点 *J*

聚集系数的变化。对于图 3 给出的拓扑结构，按照 k-核解析的原则可以得出如下推论。

1) 由于 k-核中节点 J 的度值为 $d_{J(k)}$，因此，k-核中节点 I 的度值 $d_{I(k)}$最大为 $d_{J(k)}-2$，因为若 $d_{I(k)}=d_{J(k)}-1$，则当 I 被去除后节点 J 的度值也将变成 $d_{J(k)}-1$，依照 k-核解析定义，J 节点也将被去除，这与原设定不符。

2) 当 k-核内节点 I 被去除后，(k+1)-核中节点 J 邻居节点连边数目相对于 k-核中节点 J 邻居节点连边数最多减少 $d_{J(k)}-3$ 条，即 $E_{J(k)}-E_{J(k+1)}<= d_{J(k)}-3$，因为即使节点 I 在 k-核内取最大度值 $d_{J(k)}-2$,节点 J 邻居节点连边数目中 I 节点的所占据的连边数目应为 I 的度值减去 I 与 J 之间的一条连边，即为 $d_{J(k)}-2-1$。

3) 当 k-核内节点 I 被去除后，k+1-核内每个节点度值最小为 $d_{I(k)}-1$，因为若某个节点度值为 $d_{I(k)}-2$，则该节点将与节点 I 一起被去除。

图 4 中给出了 k-核、(k+1)-核中聚集系数的函数直线，可以看出，$E_{J(k)}$与 $E_{J(k+1)}$的差值若可以取任意值，则 $CC_{J(k)}$与 $CC_{J(k+1)}$的大小无法确定。但是通过上述结论可知 $E_{J(k)}$与 $E_{J(k+1)}$的差值最高为 $d_{J(k)}-3$，在这个限制下 $CC_{J(k)}$与 $CC_{J(k+1)}$存在何种大小关系呢？

从图 4 中可以看出在给定某个 $E_{J(k)}$值的前提下，随着 $E_{J(k)}$与 $E_{J(k+1)}$的差值由零逐步变大，将会依次出现 $CC_{J(k)}<CC_{J(k+1)}$、$CC_{J(k)}=CC_{J(k+1)}$ 和 $CC_{J(k)}>CC_{J(k+1)}$的关系。$E_{J(k)}$与 $E_{J(k+1)}$的差值越小，则 $CC_{J(k)}>CC_{J(k+1)}$越不可能出现。因此，若当 $E_{J(k)}-E_{J(k+1)}= d_{J(k)}-3$ 时，$CC_{J(k)}<CC_{J(k+1)}$，则可以得出结论：无论 $E_{J(k)}$取何值，k+1-核对应的聚集系数将大于 k 核对应的聚集系数。当 $E_{J(k)}- E_{J(k+1)}= d_{J(k)}-3$ 时，假设式(5)成立

$$CC_{J(k)}=\frac{2E_{J(k)}}{l_{J(k)}(l_{J(k)}-1)}<CC_{J(k+1)}=\frac{2E_{J(k+1)}}{l_{J(k+1)}(l_{J(k+1)}-1)} \quad (5)$$

则有

$$\begin{aligned}&\frac{2E_{J(k)}}{l_{J(k)}(l_{J(k)}-1)}<\frac{2E_{J(k+1)}}{l_{J(k+1)}(l_{J(k+1)}-1)}\Rightarrow\\&\frac{E_{J(k+1)}+d_{J(k)}-3}{l_{J(k)}(l_{J(k)}-1)}<\frac{E_{J(k+1)}}{(l_{J(k)}-1)(l_{J(k)}-2)}\Rightarrow\\&\frac{E_{J(k+1)}+d_{J(k)}-3}{l_{J(k)}}<\frac{E_{J(k+1)}}{l_{J(k)}-2}\end{aligned} \quad (6)$$

又因为图 3 中节点 J 的度值就是节点 J 邻居节点的个数，因此 $d_{J(k)}=l_{J(k)}$，所以式(6)可以演化为

$$\begin{aligned}&\frac{E_{J(k+1)}+d_{J(k)}-3}{l_{J(k)}}<\frac{E_{J(k+1)}}{l_{J(k)}-2}\Rightarrow\\&\frac{E_{J(k+1)}+d_{J(k)}-3}{d_{J(k)}}<\frac{E_{J(k+1)}}{d_{J(k)}-2}\Rightarrow\\&E_{J(k+1)}>\frac{(d_{J(k)}-2)(d_{J(k)}-3)}{2}\end{aligned} \quad (7)$$

由于节点 I 取得度值为 $d_{J(k)}-2$，所以在 k+1-核内每个节点的度值最小为 $d_{J(k)}-1$，去除与节点 J 的连边。则在 k+1-核内，节点 J 的邻居节点共有 $d_{J(k)}-1$ 个，每个节点度值最小为 $d_{J(k)}-2$。通过拓扑学不难得出，这 $d_{J(k)}-1$ 个度值最小为 $d_{J(k)}-2$ 的邻居节点，构成的拓扑就是全连通拓扑结构，此时的 $E_{J(k+1)}$为

$$E_{J(k+1)}=\frac{(d_{J(k)}-1)(d_{J(k)}-2)}{2}>\frac{(d_{J(k)}-2)(d_{J(k)}-3)}{2} \quad (8)$$

式(8)证明了式(7)的成立，因此假设成立，即 $CC_{J(k)}<CC_{J(k+1)}$。

通过上述证明可以得出 2 点结论。

1) 在 k-核解析过程中若被去除节点 I 的度值为 $d_{J(k)}-2$，则 k+1 核中节点 J 的邻居节点间将为全连通结构。

2) 在给定拓扑结构中进行 k-核解析使拓扑结构由 k-核变为 k+1-核，当去除一个低度值节点 I 时，若 I 节点的去除对原拓扑结构中节点 J 的聚集系数产生影响，且节点 J 保留至高核中时，则节点 J 在 k+1-核中对应的聚集系数大于 k-核中对应的聚集系数。即随着 k-核解析的进行，高核节点的聚集系数保持不变或者增高。

由上述结论不难看出，随着低核节点的逐个去除，一方面导致高核中节点数目的降低，另一方面保留在高核中各个节点的聚集系数只增不减，因此高核拓扑结构最终对应的网络聚集系数将会增加，即命题 1 成立。至此，关于命题 1 的证明结束。

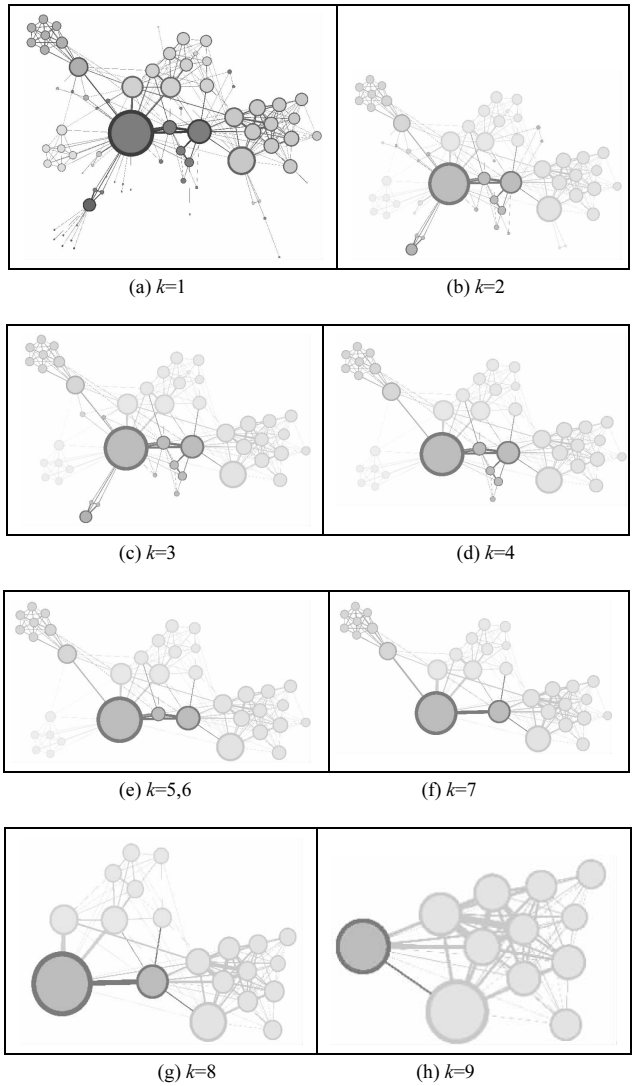
4　实验及仿真

为了检验命题 1 的正确性，本文设计了一个实验，首先实现了 Les Miserables 网络生成方法[12]，生成了一个复杂网络，构建了网络拓扑结构布局，然后借助于 Gephi 复杂网络性能分析平台，逐步进行 k-核解析，统计每一个 k 值对应的网络聚集系数。若随着 k 值的增加，网络聚集系数呈增长趋势，则能够验证命题 1 的正确性。实验相关的设置如表 1 所示。

表 1 网络仿真参数设置

项目	数值
图类型	无向图
节点数	77
边数	254
动态图	No
分层图	No
自动标尺	Yes
创建丢失节点	Yes

图 5 显示了表 1 仿真环境下，不同 k 值设定下的网络拓扑结构分解图，可以看出，随着 k 值的增加，网络中节点越来越少，当 k=10 时，网络消失，因此表 1 给出的网络拓扑最高核为 9。此外当 k=5 与 k=6 时，网络拓扑结构没有发生变化，因此当 k=5 时的网络拓扑结构中所有节点度值均大于 6。



(a) k=1 (b) k=2

(c) k=3 (d) k=4

(e) k=5,6 (f) k=7

(g) k=8 (h) k=9

图 5 k–1 核解析分解

各个 k 值对应的网络聚集系数如表 2 所示。

表 2 k 值与网络聚集系数的对照

k 值	聚集系数
1	0.736
2	0.777
3	0.822
4	0.826
5,6	0.826
7	0.827
8	0.827
9	0.952

从表 2 可以看出，随着 k-核的不断解析、k 值的不断增加，网络聚集系数呈现逐步增加的趋势，该仿真测试结果与定理 1 相吻合。

5 结束语

本文主要对 k-核解析与网络聚集系数之间的关联进行了论证研究。k-核解析是一种复杂图分解的常见方法，但是随着 k-核的不断分解，高核结构所体现出的特性到底如何变化是一项研究空白，本文针对网络聚集系数这一特性，展开了研究。通过理论推导与证明，明确了 k-核分解与网络聚集系数之间的关联，即越高核对应的网络聚集系数越高，在此基础上设计了实验检验理论证明。本文得出的结论可以与现有宏观统计分析方法相结合，为复杂网络应用提供相应的理论基础，例如重要节点或者结构发掘研究、网络抗毁性研究等。以文中实验为例，当 9-核网络对应的聚集系数为 0.952，这表明该网络中的小世界性很高，具有高传播性和高抗毁性，对病毒预防领域中，该结构的危险性最高，通过本文后续研究能够找出最优结构调整策略，在微小删边代价下重构 9-核，将能大大提高网络病毒免疫能力。

在后续工作中，本文将重点研究网络节点、边的价值属性，即当删除/添加一个节点或者一条边对网络的聚集系数以及其他特征参数所产生的影响。k-核解析在某种意义上也是一种节点、边的删除行为，但是本文只是通过推导证明了 k-核分解会造成网络聚集系数的增加，但是并没有给出具体增加的幅度。通过后续工作的研究，能够为 k-核解析提供一种全新的量化视角。

参考文献：

[1] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用[M]. 北京：清华大学出版社, 2006.49-70.
WANG X F, LI X, CHEN G R. Theory and Applications of Complex Network[M]. Beijing: Tsinghua University Press, 2006.49-70.

[2] SONG C, HAVLIN S, MAKSE H A. Origins of fractality in the growth of complex network[J]. Nature Physics, 2006, 2(4): 275-281.

[3] GOH K I, SALVI G, KAHNG B, *et al*. Skeleton and fractal scaling in complex networks[J]. Physical Review Letters, 2006, 96: 018701.

[4] GUO Q Z, *et al*. Exploring the local connectivity preference in Internet AS level topology[J]. Piscataway, 2007, 13(1): 6439-6445.

[5] ZEGURA E W, CALVERT K L, DONAHOO M I. A quantitative comparison of graph-based models for internet topology[J]. IEEE/ACM Trans on Networking, 1997, 5(6):770-783.

[6] ONNELA, J. SARAMAKI, HYVONEN J. Structure and tie strengths in mobile commutation networks[J]. PNAS, 2007,104(18): 7332- 7336.

[7] BARCELÓ J M, NIETO-HIPÓLITO J I, GARCIA-VIDAL J. Study of Internet autonomous system interconnectivity from BGP routing tables[J]. Computer Networks, 2004, 45(3):333-344.

[8] DIMITROPOULOS X, KRIOUKOV D, FOMENKOV M, *et al*. AS relationships: inference and validation[J]. SIGCOMM Computer Communications Review, 2007, 37(1):29-40.

[9] PARK S T, PENNOCK D M, GILES C L. Comparing static and dynamic measurements and models of the Internet's topology[A]. Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies[C]. 2004.1616-1627.

[10] ZHOU S, MONDRAGON R J. The rich-club phenomenon in the Internet topology[J]. IEEE Communication Letters, 2004, 8(3): 180-182.

[11] ZHOU S, MONDRAGON R J. Structural constraints in complex networks[J]. New Journal of Physics, 2007, 9(172):1-11.

[12] KNUTH D E. The Stanford GraphBase: A Platform for Combinatorial Computing[M]. Addison-Wesley, Reading, MA, 1993.

作者简介：

刘君（1982-），女，山东潍坊人，博士，东北大学讲师，主要研究方向为复杂网络、分布式计算等。

乔建忠（1964-），男，辽宁沈阳人，博士，东北大学教授、博士生导师，主要研究方向为网格计算、分布式计算等。