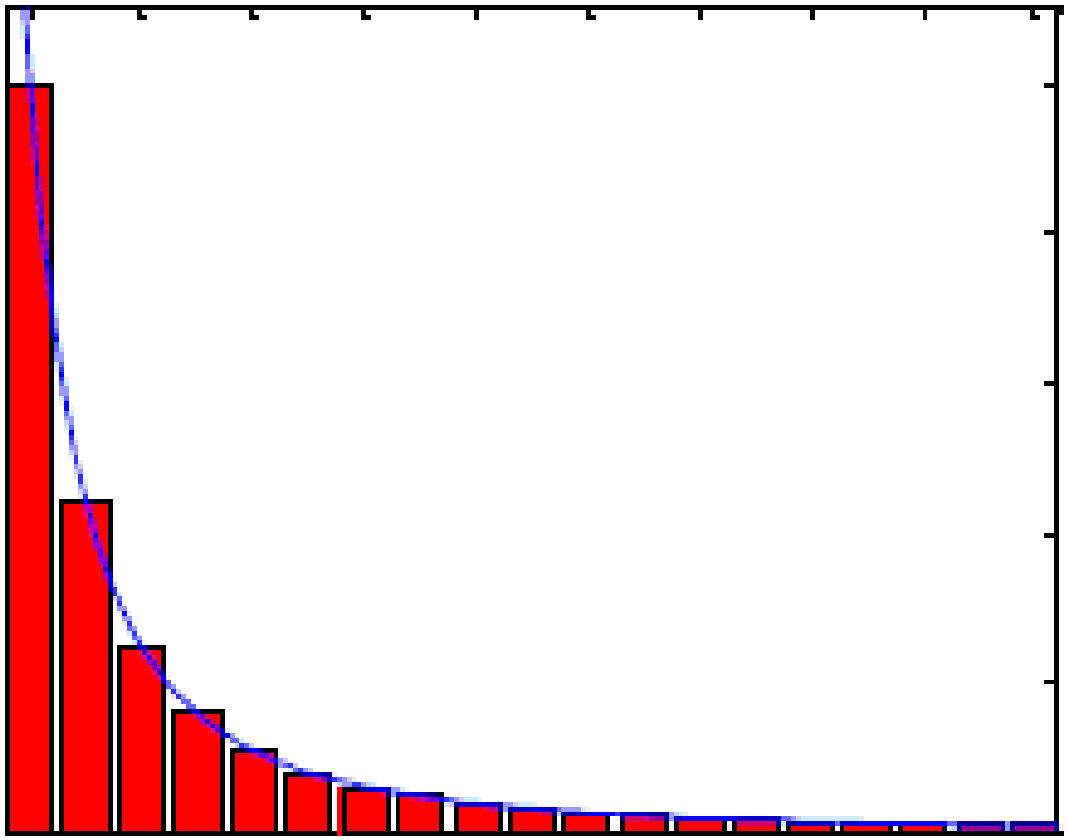
Universal distribution of component frequencies in biological and technological systems

报告人：程华

2013年11月13日

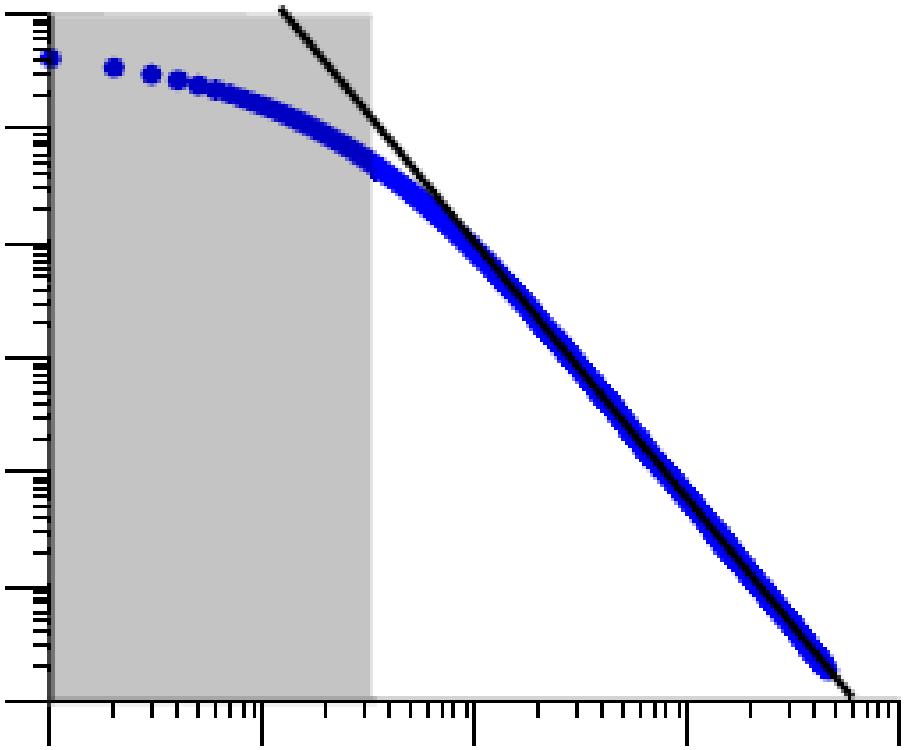
引言

- **无尺度网络scale-free：**是指在某一复杂的系统中，大部分节点只有少数几个连结，而某些节点却拥有与其他节点的大量连结。这些具有大量连结的节点称为“集散节点”，所拥有的连结可能高达数百、数千甚至数百万。这一特性说明该网络是无尺度的，因此，凡具有这一特性的网络都是无尺度网络。



- 幂律分布（power-law distribution）

- 1932年，哈佛大学的语言学专家Zipf在研究英文单词出现的频率时，发现如果把单词出现的频率按由大到小的顺序排列，则每个单词出现的频率与它的名次的常数次幂存在简单的反比关系：$P(r) \sim r^{\gamma(-\alpha)}$，这种分布就称为Zipf定律。



- The first category invokes random multiplicative processes recently exemplified by the preferential attachment model of growing networks

- The second category of models invokes heterogeneity of functional roles of individual components

材料和方法

- 529 bacterial genomes and 44,283 prokaryotic orthologous gene families
- 1,832 reactions/enzymes connected to each other by 3,118 direct and 49,168 direct+indirect dependencies.
- 192,392 packages on 2,047,796 computers
- 33,473 packages, 157,667 direct, and 2,439,011 total dependency relations

材料和方法

- http://tuvalu.santafe.edu/~aaronc/powerlaws.

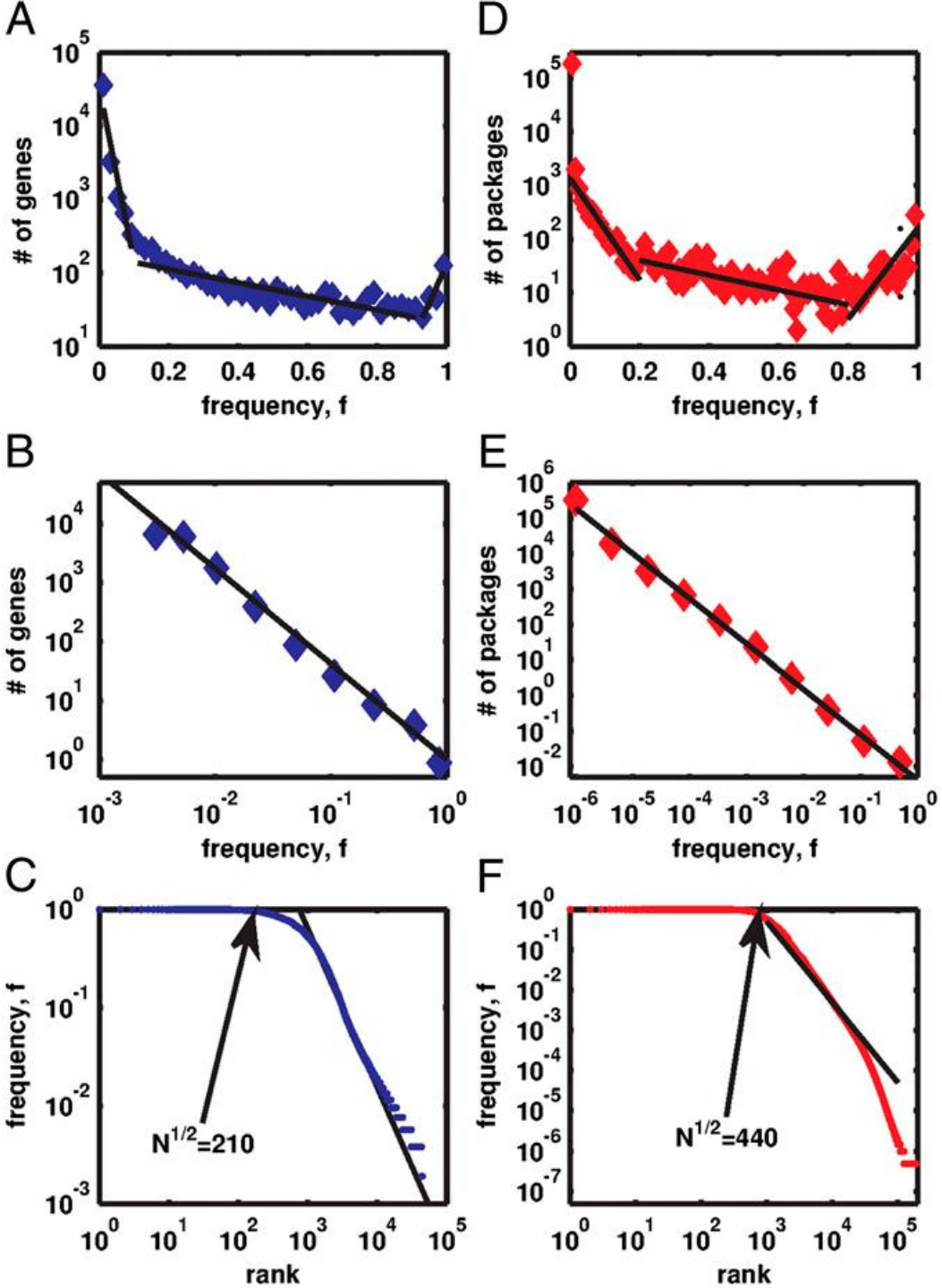
Fitting a power-law distribution

This function implements both the discrete and continuous maximum likelihood estimators for fitting the power-law distribution to data, along with the goodness-of-fit based approach to estimating the lower cutoff for the scaling region. Usage information is included in the file; type 'help plfit' at the Matlab prompt for more information.

plfit.m (Matlab, by Aaron Clauset)
plfit.r (R, by Laurent Dubroca)
plfit.py (Python, by **Adam Ginsburg**)
plfit.c (C++, by Wim Otte; includes plvar.c)
plfit.c (C++, by Tamas Nepusz)
plfit.py (Python, by Joel Ornstein)

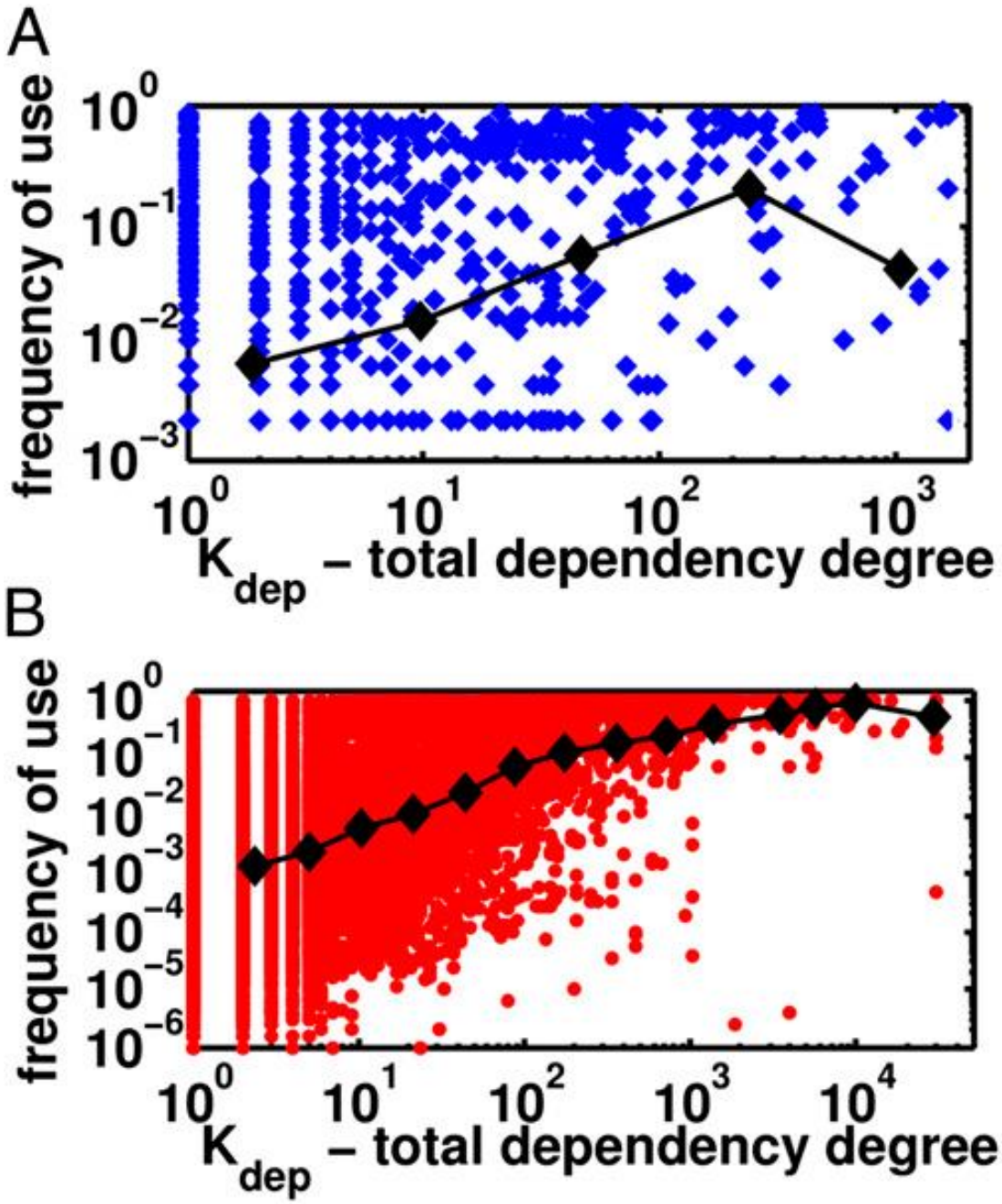
试验结果

1.组件频率的分布



2.组件频率与依赖度正相关

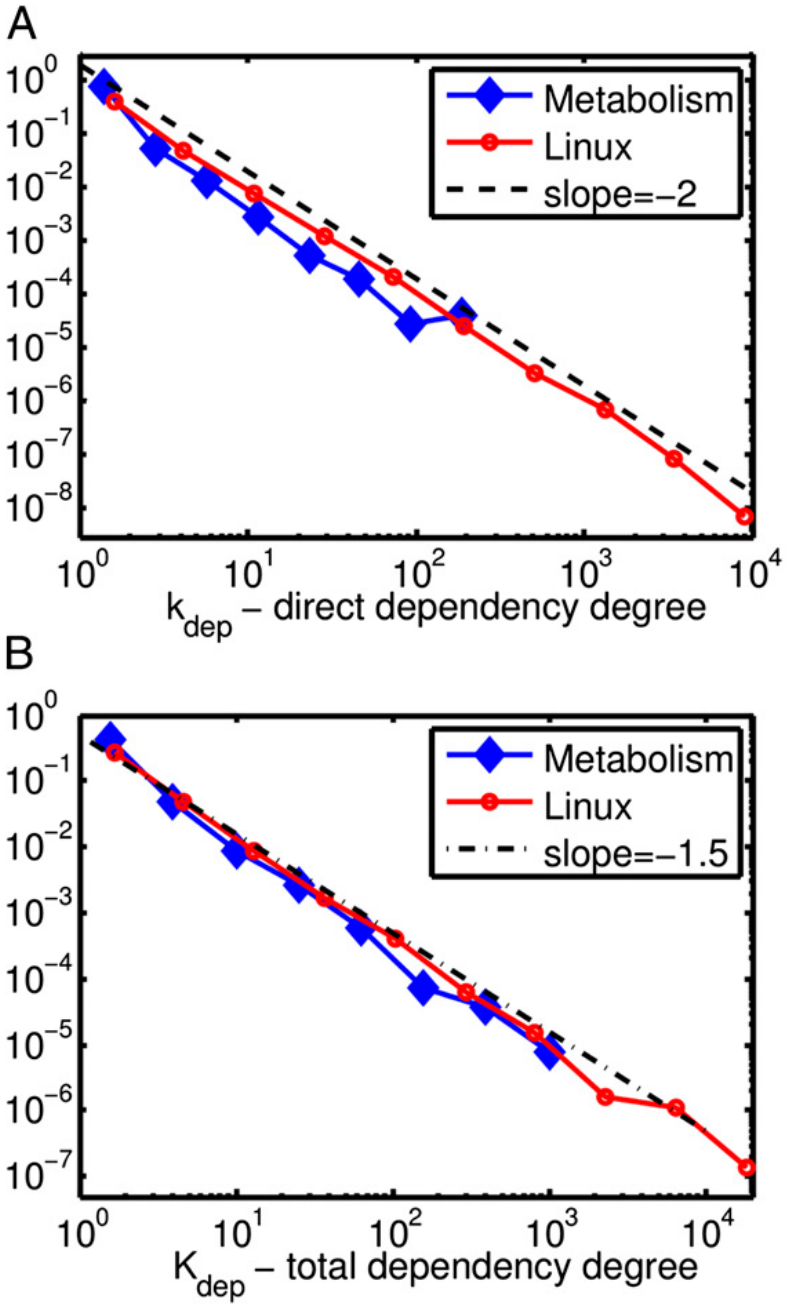
- $k_{dep}(i)$ counts the packages that require installation of the package i at the first step of this multistep process
- $K_{dep}(i)$ counts the packages that do so at any step
- $k_{dep}(i)$ counts enzymes located one step below (or above) it in this hierarchy
- $K_{dep}(i)$ of the enzyme i is given by the total number of enzymes in this minimal pathway located downstream from it for anabolic enzymes (or upstream from it for catabolic enzymes)



- Fig. 2. Components' frequencies f (y axis) are positively correlated with their total (direct + indirect) dependency degrees Kdep (x axis) for both metabolic enzymes (A) (Spearman's rs = 0.30) and Linux packages (B) (Spearman's rs = 0.47). The black lines and symbols show the geometric averages of f in each logarithmic bin of Kdep.

3.依赖度符合幂律分布

- Fig. 3. Probability distributions of direct (kdep; A) and total (Kdep; B) dependency degrees for metabolic enzymes (blue diamonds) and Linux packages (red circles). Power-law fits to direct degree cumulative distribution give −2.08 for metabolic enzymes and −1.91 for Linux packages, and are both consistent with the −2.0 scaling law (solid line in A). Power-law fits to direct degree cumulative distribution give −1.5 for metabolic enzymes and −1.56 for Linux packages, consistent with the mathematically derived −1.5 scaling (solid line in B).



讨论

- one is optimized by nature over billions of years of evolution

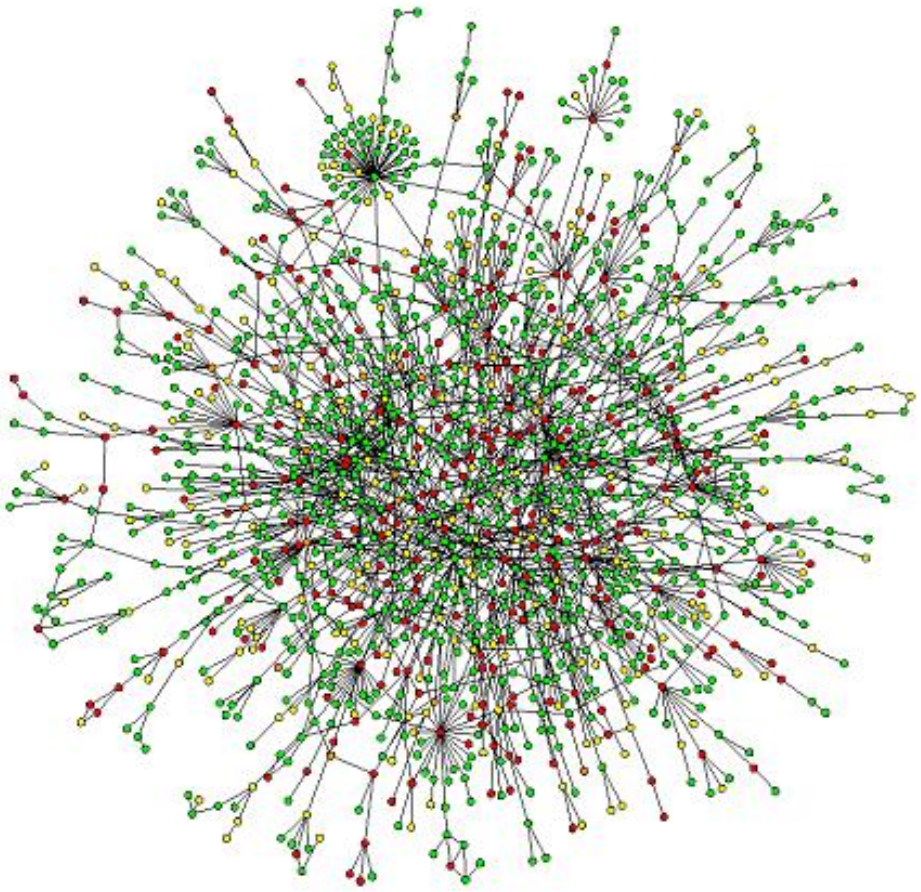
- The other is designed by a distributed population of human software engineers over the past several decades

讨论

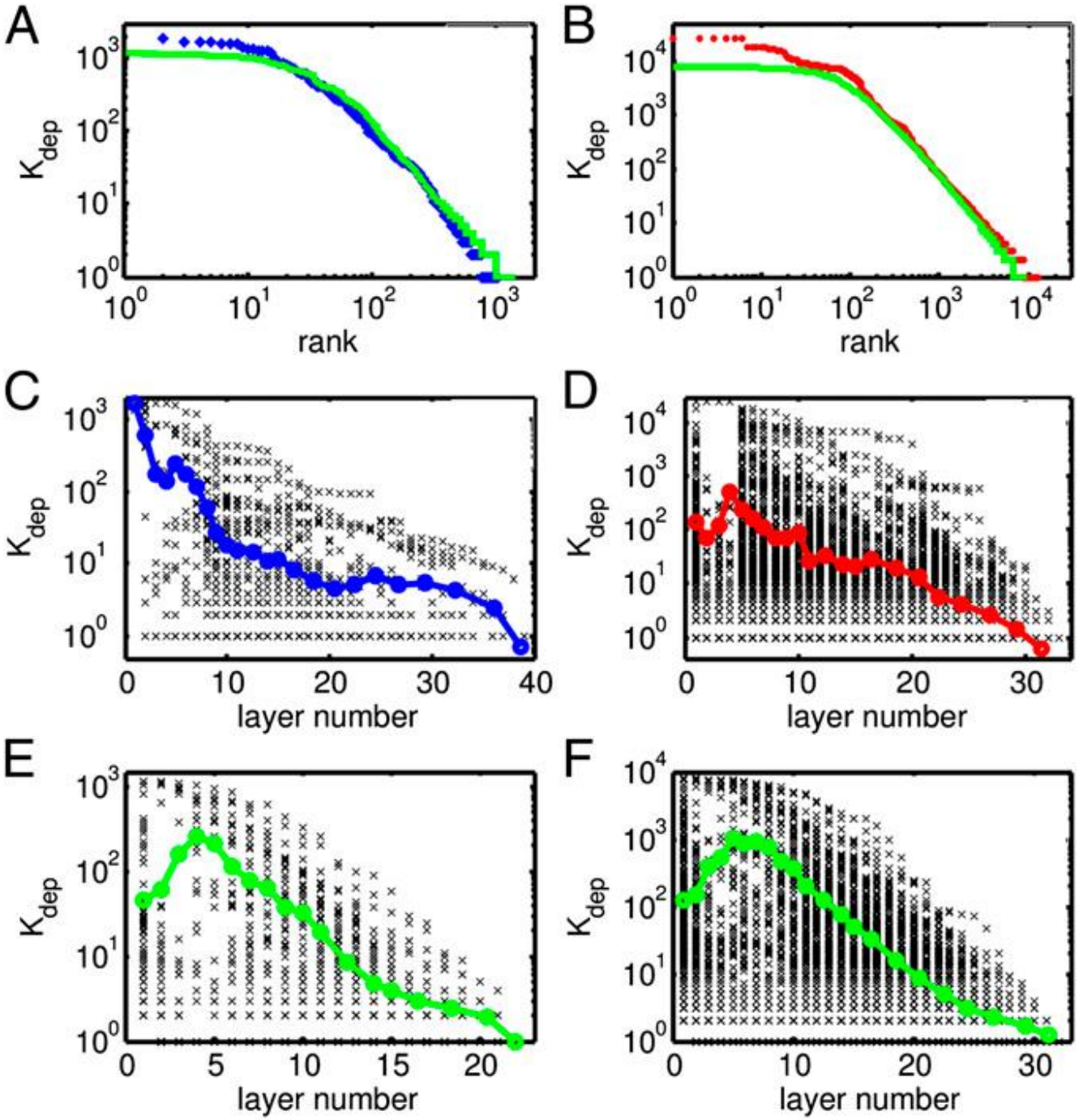
- In a tree, each component directly depends on one, and only one, downstream component.

讨论

- In real-life networks this number, D, is certainly larger than one; it varies from component to component, but averages ~2 for both metabolic networks and Linux packages.



Yeast protein interaction network


A
B
C
D
E
F
K_{dep}
rank
layer number
10^0
10^1
10^2
10^3
10^4
0
5
10
15
20
30
40

- An important caveat in applying the $N_c = N^{1/2}$ relationship is that N counts only those components that are directly or indirectly connected to the core by the functional dependency network.

- To reconcile the apparent stability of *Nc* with unlimited growth of N, one recalls that continuing expansion of *N* is caused by either nonfunctional (prophages or transposable elements) or extremely niche-specific gene families—both are likely to be disconnected from the core and hence will not contribute to growth of *Nc*.

- A more systematic analysis of similarities and differences between different versions of biological and technological complex systems will have to await future studies.

祝大家学习愉快！