Notes on the Algorithm for Calculating Betweenness *

ZHOU Tao(周涛)[1**], LIU Jian-Guo(刘建国)[2], WANG Bing-Hong(汪秉宏)[1***]

[1]*Department of Modern Physics and Nonlinear Science Center, University of Science and Technology of China, Hefei 230026*

[2]*Institute of System Engineering, Dalian University of Technology, Dalian 116023*

(Received 9 May 2006)

We investigate a common used algorithm [Phys. Rev. E **64** *(2001) 016132] to calculate the betweenness centrality for all vertices. The inaccurateness of that algorithm is pointed out and a corrected algorithm, also with O(MN) time complexity, is given. In addition, the comparison of calculating results for these two algorithm aiming at the protein interaction network of yeast is shown.*

PACS: 89. 75. Hc, 89. 65. −s, 89. 70. +c, 01. 30. −y

Betweenness centrality, also called the load or betweenness for simplicity, is a quite useful measure in the network analysis. This concept was first proposed by Anthonisse[1] and Freeman[2] and was introduced to the physics community by Newman.[3] The betweenness of a node v is defined as

$$B(v) := \sum_{s\neq t, s\neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}, \qquad (1)$$

where $\sigma_{st}(v)$ is the number of shortest paths going from s to t passing through v, and σ_{st} is the total number of shortest paths going from s to t. The end points of each path is counted as a part of the path.[3] Newman proposed a very fast algorithm taking only O(MN) time to calculate the betweenness of all vertices,[3] where M and N denote the number of edges and vertices, respectively. The whole algorithm processes are described as follows: (1) Calculate the distance from a vertex s to every other vertex by using breadth-first search. (2) A variable b_v^s, taking the initial value 1, is assigned to each vertex v. (3) Going through the vertices v in order of their distance from s, starting from the farthest, the value of b_v^s is added to corresponding variable on the predecessor vertex of v. If v has more than one predecessor, then b_v^s is divided **equally** between them. (4) Go through all vertices in this fashion and record the value b_v^s for each v. Repeating the entire calculation for every vertex s, the betweenness for each vertex v is obtained as

$$B(v) = \sum_s b_v^s. \qquad (2)$$

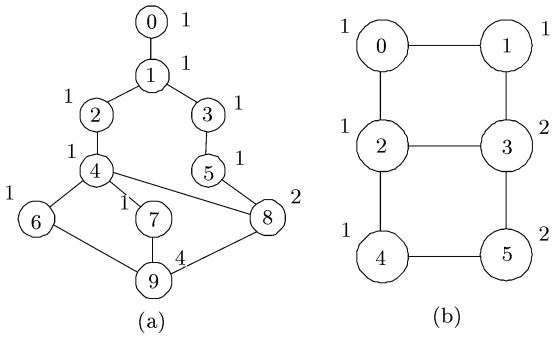

Fig. 1. Two examples used to illuminate the difference between Newman's and the corrected algorithms. (a) The copy from Ref. [2] also has been used as a sketch map for Newman's algorithm. (b) The minimal network that can illuminate the difference. The hollow circles represent the vertices and the solid lines represent the edges. Each vertex is marked with a natural number inside the corresponding circle, and the number beside each vertex v is σ_{0v}.

Table 1. Calculation results of Fig. 1(a)

Vertices	0	1	2	3	4	5	6	7	8	9
Newman's	9	$34\frac{5}{6}$	$28\frac{1}{6}$	$22\frac{1}{2}$	$29\frac{1}{3}$	$21\frac{2}{3}$	$14\frac{1}{4}$	$14\frac{1}{4}$	$21\frac{5}{6}$	$24\frac{1}{6}$
Corrected	9	$34\frac{1}{3}$	$28\frac{1}{3}$	$21\frac{2}{3}$	30	21	$14\frac{2}{3}$	$14\frac{2}{3}$	$21\frac{2}{3}$	24

Table 2. Calculation results of Fig. 1(b)

Vertices	0	1	2	3	4	5
Newman's	$6\frac{3}{4}$	$6\frac{3}{4}$	$11\frac{1}{2}$	$11\frac{1}{2}$	$6\frac{3}{4}$	$6\frac{3}{4}$
Corrected	$6\frac{2}{3}$	$6\frac{2}{3}$	$11\frac{2}{3}$	$11\frac{2}{3}$	$6\frac{2}{3}$	$6\frac{2}{3}$

* Supported by the National Natural Science Foundation of China under Grant Nos 10472116, 10532060, 10547004, 70471033 and 70571074, the Special Research Funds for Theoretical Physics Frontier Problems (NSFC No A0524701), and the President Fund of Chinese Academy of Science.

** Email: zhutou@ustc.edu

*** Email: bhwang@ustc.edu.cn

To a vertex v's betweenness $B(v)$, the contributions of its predecessors are not equal, and it is not proper to divide b_v^s equally between them. Clearly, if the vertex v has n predecessors labelled as $u_1, u_2, \cdots, u_n$ and σ_{sv} different shortest paths to vertex s, then we have

$$\sigma_{sv} = \sum_{i=1}^{n} \sigma_{su_i}. \tag{3}$$

The different shortest paths from s to v are divided into n sets $G_1, G_2, \cdots, G_n$. The number of elements in G_i, that is also the number of different shortest paths from s to u_i, gives expression to the contribution of the predecessor u_i to v's betweenness. Therefore, the vertex v's betweenness, induced by the given source s, should be divided proportionally to σ_{su_i} rather than equally between its predecessors. The corrected algorithm is as follows: (1) Calculate the distance from a vertex s to every other vertex by using breadth-first search, taking time $\mathrm{O}(M)$. (2) Calculate the number of shortest paths from vertex s to every other vertex by using dynamic programming[4], taking time $O(M)$ too. The processes are as follows. (i) Assign $\sigma_{ss} = 0$. (ii) If all the vertices of distance $d(d \geq 0)$ is assigned (Note that the distance from s to s is zero), then for each vertex v whose distance is $d+1$, assign $\sigma_{sv} = \sum_u \sigma_{su}$ where u runs over all v's predecessors. (iii) Repeat from step (i) until there are no unassigned vertices left. (3) A variable β_v^s, taking the initial value 1, is assigned to each vertex v. (4) Going through the vertices v in the order of their distance from s, starting from the farthest, the value of β_v^s is added to corresponding variable on the predecessor vertex of v. If v has more than one predecessor $u_1, u_2, \cdots, u_n$, β_v^s is multiplied by $\sigma_{su_i}/\sigma_{sv}$ and then added to σ_{su_i}. (5) Go through all vertices in this fashion and records the value β_v^s for each v. Repeat the entire calculation for every vertex s, the betweenness for each vertex v is obtained as

$$B(v) = \sum_s \beta_v^s. \tag{4}$$

Clearly, the time complexity of the corrected algorithm is also $\mathrm{O}(MN)$. In addition, one should pay attention to a more universal algorithm proposed by Brandes,[5] which can be used to calculate all the kinds of centrality based on shortest-paths counting for both unweighted and weighted networks.

These two algorithms, Newman's and the corrected one, will give the same result if the network has a tree structure. However, when the loops appear in the networks, the diversity between them can be observed. Figure 1 exhibits two examples, the first one is copied from Ref. [2], and the second is the minimal network that can illuminate the difference between Newman's and the corrected algorithms. The comparisons between these two algorithms are performed in Tables 1 and 2. The two algorithms produce different results even for networks of very few vertices.

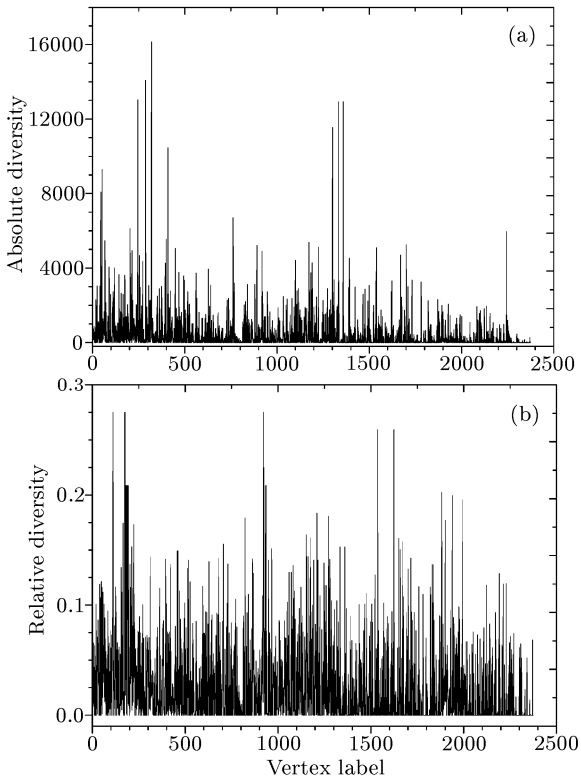

Fig. 2. The comparison between Newman's and the corrected algorithms on the protein interaction network of yeast. Here (a) the absolute diversity and (b) relative diversity between Newman's and the accurate results are presented.

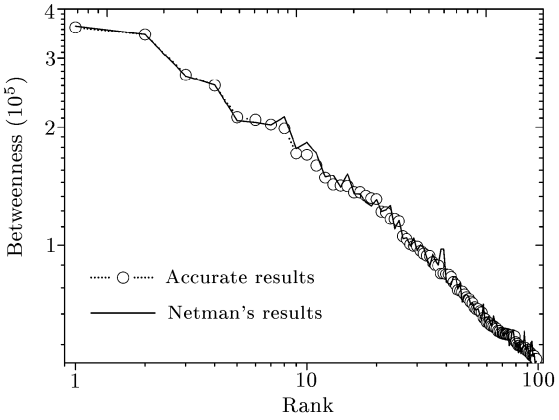

Fig 3. The Zipf plot of the 100 vertices with highest betweenness of the protein interaction network.

In addition, we compare the performances of these two algorithms on the protein interaction network of yeast.[6] This network has 2617 vertices, but only its maximal component containing 2375 vertices is taken into account. Figures 2(a) and 2(b) report the abso-

lute diversity and relative diversity between Newman's and the accurate (obtained from the corrected algorithm) results, respectively. The departure is distinct and cannot be neglected. Fortunately, the statistical features may be similar. Although the details of the Zipf plot[7] of the top-100 vertices are not the same, both the two curves obey power-law form with almost the same exponent (see Fig. 3 for details). We also have checked that the scaling law[8,9] of betweenness distribution in Barabási-Albert networks[10] is kept, while the power-law exponents are slightly changed.

The measure of betweenness is now widely used to detect communities/modules structures[11,12] and to analyse dynamics upon networks. Since the statistical characters of betweenness distributions obtained by Newman's and the corrected algorithm are almost the same, some researchers may have found the difference between these two algorithm but have not paid attention to it. However, many previous works have demonstrated that betweennesses of a few nodes rather than the overall betweenness distribution may sometimes determine the key features of dynamic behaviour on networks. Examples are numerous: these include the network traffics,[13–15] the synchronization,[16–19] the cascading dynamics,[20,21] and so on. In Fig. 2(b), one can find that for many nodes, the relative diversities between those two algorithms exceed 10%, and even nearly 30% for a few nodes. Therefore, the difference cannot be neglected especially in analysing the networks dynamics.

Although Newman's algorithm does not agree with the definition of betweenness,[3] it may be more practical especially for the large-scale communication systems wherein the routers do not know how many shortest paths there are to the destination. Even if they can save the information of weights of all the successors, to implement the biased choices may bring additional costs in economy and technique. Hence just to choose with equal probability at each branch point may be more natural, which is in accordance with Newman's algorithm.

References

[1] Anthonisse J M 1971 *Technical Report BN 9/71* (Amsterdam: Stichting Mathematisch Centrum)
[2] Freeman L C 1977 *Sociometry* **40** 35
[3] Newman M E J 2001 *Phys. Rev.* E **64** 016132
[4] Bellman R E and Dreyfus S E 1962 *Applied Dynamic Programming* (Princeton, NJ: Princeton University Press)
[5] Brandes U 2001 *J. Math. Soc. Am.* **25** 163
[6] Jeong H, Mason S, Barabási A L and Oltvai Z N 2001 *Nature* **411** 41
[7] Zipf G K 1949 *Human Behavior and the Principal of Least Effort* (Cambridge, MA: Addison-Wesley)
[8] Goh K I, Kahng B and Kim D 2001 *Phys. Rev. Lett.* **87** 278701
[9] Goh K I, Oh E, Jeong H, Kahng B and Kim D 2002 *Proc. Natl. Acad. Sci. U.S.A.* **99** 12583
[10] Barabási A L and Albert R 1999 *Science* **286** 509
[11] Girvan M and Newman M E J 2002 *Proc. Natl. Acad. Sci. U.S.A.* **99** 7821
[12] Newman M E J and Girvan M 2004 *Phys. Rev.* E **69** 026113
[13] Guimerá R, Díaz-Guilera A, Vega-Redondo F, Cabrales A and Arenas A 2002 *Phys. Rev. Lett.* **89** 248701
[14] Zhao L, Lai Y C, Park K and Ye N 2005 *Phys. Rev.* E **71** 026125
[15] Yan G, Zhou T, Hu B, Fu Z Q and Wang B H 2006 *Phys. Rev.* E **73** 046108
[16] Nishikawa T, Motter A E, Lai Y C and Hoppensteadt F C 2003 *Phys. Rev. Lett.* **91** 014101
[17] Hong H, Kim B J, Choi M Y and Park H 2004 *Phys. Rev.* E **69** 067105
[18] Zhao M, Zhou T, Wang B H and Wang W X 2005 *Phys. Rev.* E **72** 057102
[19] Zhou T, Zhao M and Wang B H 2006 *Phys. Rev.* E **73** 037101
[20] Motter A E 2004 *Phys. Rev. Lett.* **93** 098701
[21] Zhou T and Wang B H 2005 *Chin. Phys. Lett.* **22** 1072