SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Mining the rank of universities with the encyclopedia Wikipedia

Zongjian LI[1], Cong LI[1,2*] & Xiang LI[1,2]

[1] *Adaptive Networks and Control Lab,Department of Electronic Engineering, Fudan University, Shanghai* 200433*,P.R.China;*
[2] *Research Center of Smart Networks and Systems, School of Information Science Engineering,*
Fudan Univeristy,Shanghai 200433*, P.R.China*

Appendix A Fetching the network of *Wikipedia* articles

The web crawler fetches the HTML file, which contains the hyperlinks of each *Wikipedia* article. The hyperlink $L_{i,j}$ is a directed link, which directs from a *Wikipedia* article i to another *Wikipedia* article j.The obtained *Wikipedia* data can be mapped into a raw *Wikipedia* network, where the *Wikipedia* articles are vertices, and the hyperlink between a pair of *Wikipedia* articles is the edge (link) between the corresponding two vertices. Note that the article “Main Page” and article “UTC”, which are linked from all articles, are ignored in this work. Algorithm A1 illustrates how the raw data is collected and stored into the database.
The network generated from the raw data contains multiple nodes, since articles may have aliases. For instance, the article titled with “Fudan University” may be corresponding to the node “Fudan” with the URL (en.wikipedia.org\wiki\Fudan) or the node “Fudan University” with URL (en.wikipedia.org\wiki\Fudan University). In this work, we only consider the simple networks, which do not contain multiple nodes. Therefore, we merge all the alias of the same article into one node and redirect all the corresponding URLs to the merged node. The newly generated network is named the *Wikipedia* article reference network (WARN). Algorithm A2 depicts the procedure of generating WARN with the network database. The process ensures that each node in the network is a unique *Wikipedia* article.

Appendix B Data Description

The geographic distribution of the 114 universities is shown in Figure B1. The 114 universities are located in 18 countries (or regions). Totally, 72 universities are located in six English-speaking regions. Thirty-eight universities are in the United States, and twenty universities are in the United Kingdom.

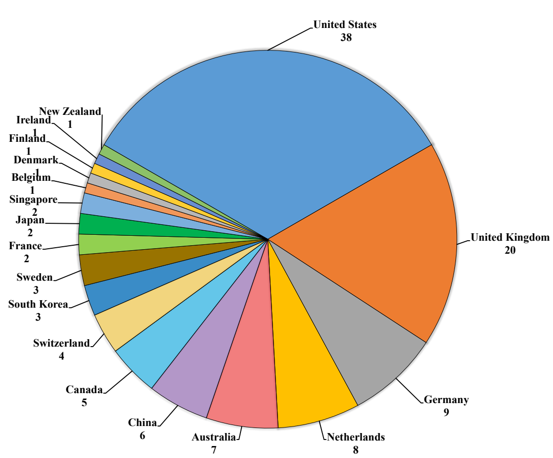


Figure B1 The geographic distribution of the 114 university subjects.

* Corresponding author (email: cong_li@fudan.edu.cn)

Algorithm A1 Fetch data using crawler program and network database.

```
Ensure:
Raw network R = (N, L);
Redirect information dictionary D;

D ← ∅
N ← {MainPage}
L ← ∅
S ← {MainPage}                                ▷ A set to store unvisited article web pages.
while S ≠ ∅ do
    p ← Pop(S)
    a ← DownloadArticle(p)
    t ← GetArticleTitleInHtml(a)
    if t ≠ p then
        D[p] ← t
    else
        for all h ∈ GetHyperLinks(a) do
            if IsHyperLinkForArticle(h) then
                Push(S,h)
                N ← N ∪ {h}
                L ← L ∪ {(p, h)}
            end if
        end for
    end if
end while
N ← N − {MainPage, UTC}
for all n ∈ N do
    L ← L − {(n, MainPage), (n, UTC)}
end for
return R, D
```

Algorithm A2 Generate WARN from the raw network in network database.

Require:

Raw network $R = (N, L)$;

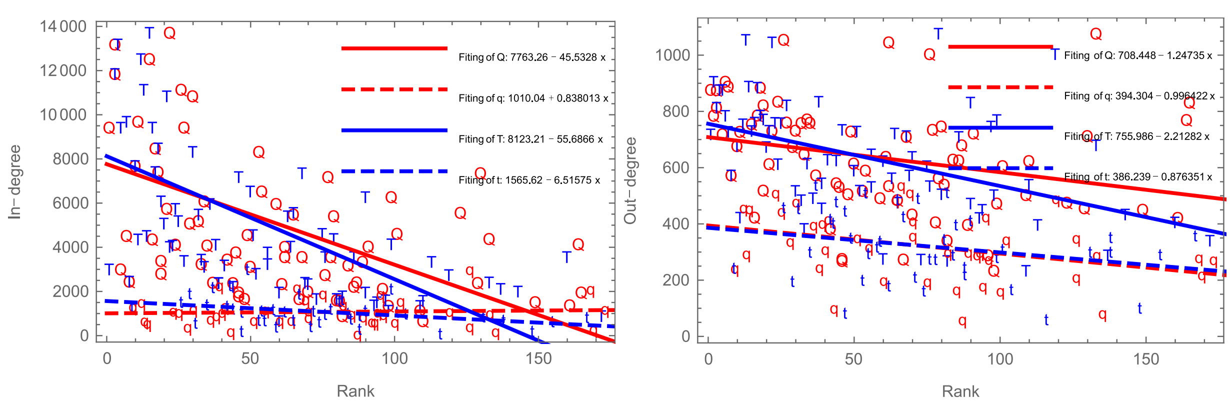
Redirect information dictionary D;

Ensure: WARN $W = (V, E)$;

```
V ← ∅
for all v ∈ N do
    if v ∉ GetKeys(D) then
        V ← V ∪ {v}
    end if
end for
E ← ∅
for all (v, w) ∈ L do
    if w ∈ GetKeys(D) then
        u ← D[w]                ▷ Get the original vertex u which has a alias w.
        E ← E ∪ {(v, u)}
    else
        E ← E ∪ {(v, w)}
    end if
end for
return W
```

Appendix C Relation between reputation criterions and the QS or THE ranking

The relation between the intuitive reputation criterion, the potential reputation criterion or the network-based criterion and the QS or THE ranking is shown in Figure C2 and Figure C1. The "T" and "t" ("Q" and "q") marks denote that their x-axis is the THE (QS) ranking. Marks in upper-case letter denote universities located in an English-speaking region, and the lower-case letter marks represent the universities in a non-English-speaking regions.



(*a*) (*b*)

Figure C1 Relation between the in-degree and out-degree of univerisity articles in *Wikipedia* and QS or THE ranking.

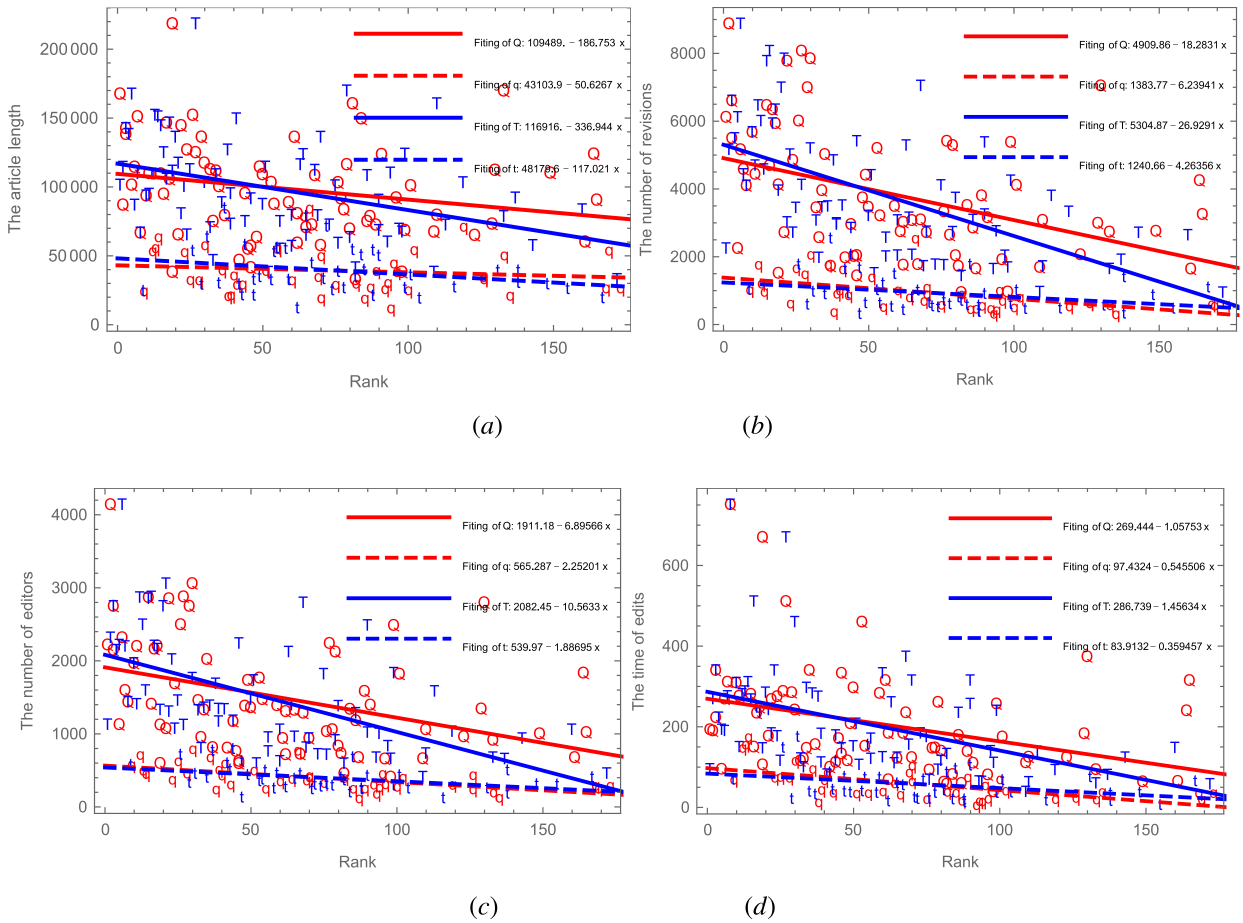


Figure C2 Relation between the intuitive reputation criterion or the potential reputation criterion of univerisities in *Wikipedia* and their QS or THE ranking.

Appendix D Algorithms for EWARS generator

We propose to generate a pre-EWARS, which includes not only the targeted articles but also the articles related to the targeted articles, as the preparation for the EWARS. Then, we perform the path length method or the local vertex connectivity method (Algorithms are introduced below) on a pre-EWARS to obtain the EWARS.

Appendix D.1 Pre-EWARS

Two parameters, *i.e.* the direction parameter and the depth parameter should be set for generating the pre-EWARS. The direction parameter is regarded as a filter through which only a specific kind of directed links pass into the pre-EWARS from the WARN. The depth parameter is the recursive depth that is used to limit the data mining depth. Algorithm D1 formulates a procedure of creating a pre-EWARS. In the initial step, we set the pre-EWARS as an empty network. Then, we find the nearest neighbors along the in-coming or out-going links from the 114 university articles and add the links and articles to the pre-EWARS. Next, we try to find the neighbors of the nearest neighbors if the depth parameter is larger than 1. All articles of the pre-EWARS will be found by recursively executing the “find and add” operation, until the depth limit is satisfied. Finally, we obtain an undirected pre-EWARS by adding inverse links (see Algorithm D1)

Appendix D.2 Path length method for generating EWARS

In the path length (PL) method, we use the shortest path length to represent the relation weight between two university articles. First, we generate a pre-EWARS. Second, the shortest path lengths between any two university articles are calculated. Third, we set the shortest path lengths as the relation weights between universities in the EWARS. Algorithm D2 shows the process of generating the EWARS.

Appendix D.3 Local vertex connectivity method for generating EWARS

Traditionally, the vertex connectivity v of a network G is defined as the minimum number of nodes whose removal disconnects the network. A network with $v > 0$ is connected, and a network with $v = k$ is said to be k-connected. In this work, we propose to utilize a local vertex connectivity method to generate EWARS. Note that the local vertex connectivity of two directly connected nodes does not exist, since we cannot disconnect two directly connected nodes by removing other nodes. Hence, when we calculate the v_{ij}, we do not take the direct connection L_{ij}

Algorithm D1 Generate a pre-EWARS from given WARN.

```
Require:
    WARN W;
    Universities U;
    Edge direction d;
    Recursive depth r;
Ensure: Pre-EWARS P = (N, L);

    N ← ∅
    L ← ∅
    for all u ∈ U do
        AddVertex(u,r)
    end for
    return P
    procedure AddVertex(v, r)
        if r > 0 then
            if d = OUT then
                for all (v, w) ∈ W do
                    N ← N ∪ {w}
                    L ← L ∪ {(v, w), (w, v)}
                    AddVertex(w, r − 1)
                end for
            else
                for all (w, v) ∈ W do
                    N ← N ∪ {w}
                    L ← L ∪ {(v, w), (w, v)}
                    AddVertex(w, r − 1)
                end for
            end if
        end if
    end procedure
```

Algorithm D2 Generate a EWARS from given pre-EWARS using PL method.

Require:

Pre-EWARS P;

Universities U;

Ensure: EWARS $G = (V, E)$;

```
V ← ∅
for all u ∈ U do
    V ← V ∪ {u}
end for
E ← ∅
for all u ∈ U do
    for all v ∈ U do
        if (u, v) ∉ E and (v, u) ∉ E then
            w ← ShortestPathLength(P,u,v)          ▷ Calculate edge weight.
            E ← E ∪ {(u, v, w), (v, u, w)}
        end if
    end for
end for
return G = (V, E)
```

between two vertices i and j into consideration (see Figure D1). The local vertex connectivity (LVC) method works similarly to the PL method, that we first generate a pre-EWARS. Then, the local vertex connectivity between any two university articles are calculated as weights of the edges between universities in the EWARS. The procedure of the LVC method is described in Algorithm D3.

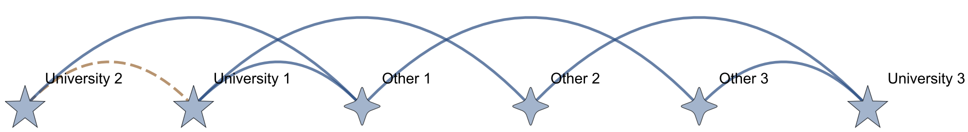


Figure D1 An example of the local vertex connectivity $v_{12} = 1$ for two directly connected nodes, university 1 and university 2.

Algorithm D3 Generate a EWARS from given pre-EWARS using LVC method.

Require:
  Pre-EWARS $P = (N, L)$;
  Universities U;
Ensure: EWARS $G = (V, E)$;

```
V ← ∅
for all u ∈ U do
    V ← V ∪ {u}
end for
E ← ∅
for all u ∈ U do
    for all v ∈ U do
        if (u, v) ∉ E and (v, u) ∉ E then
            T ← L
            L ← L − {(u, v), (v, u)}
            w ← VertexConnectivity(P,u,v)                ▹ Calculate edge weight
            L ← T
            E ← E ∪ {(u, v, w), (v, u, w)}
        end if
    end for
end for
return G = (V, E)
```