# Flow Distances on Open Flow Networks

Liangzhu Guo,[1] Xiaodan Lou,[1] Peiteng Shi,[2] Jun Wang,[2] Xiaohan Huang,[2] and Jiang Zhang[1,*]

[1] *School of Systems Sciences, Beijing Normal University, Beijing, China*
[2] *Science and Technology on Information Systems Engineering Laboratory, National University of Defence Technology, Changsha, China*

(Dated: January 27, 2015)

Open flow network is a weighted directed graph with a source and a sink, depicting flux distributions on networks in the steady state of an open flow system. Energetic food webs, economic input-output networks, and international trade networks, are open flow network models of energy flows between species, money or value flows between industrial sectors, and goods flows between countries, respectively. Flow distances (first-passage or total) between any given two nodes $i$ and $j$ are defined as the average number of transition steps of a random walker along the network from $i$ to $j$ under some conditions. They apparently deviate from the conventional random walk distance on a closed directed graph because they consider the openness of the flow network. Flow distances are explicitly expressed by underlying Markov matrix of a flow system in this paper. With this novel theoretical conception, we can visualize open flow networks, calculating centrality of each node, and clustering nodes into groups. We apply flow distances to two kinds of empirical open flow networks, including energetic food webs and economic input-output network. In energetic food webs example, we visualize the trophic level of each species and compare flow distances with other distance metrics on graph. In input-output network, we rank sectors according to their average distances away other sectors, and cluster sectors into different groups. Some other potential applications and mathematical properties are also discussed. To summarize, flow distance is a useful and powerful tool to study open flow systems.

PACS numbers: 89.90.+n

## I. INTRODUCTION

A large number of studies have proved that complex network is a powerful and useful tool to model complex systems[1–4]. However, due to the limitation of the traditional graphs for describing the complexity of the various real systems, weighted networks[5, 6], directed networks[7], bi-partite graphs[8], multiplex[9, 10], temporal networks[11] as novel extensions of the conventional graphs emerge in the past decade. Among these, open flow network is a particular kind of directed weighted network to depict open flow system.

Most complex systems are open, they exchange energy and material with their environment[12]. Energy and material flows are delivered to each unit of a system by the flow network[13, 14]. The distribution of these flows in the entire body of a system is described by directed weighted edges. Two special nodes "source" and "sink" are always added in the system to represent environment. Because the flow system considered is supposed to be in a steady state, the flow network is always balanced which means that the total inflow of each node equals to its total out flow except for the sink and the source.

Energetic food web is a typical open flow network which has been studied for several years by system ecologists. The seminal work of H.T. Odum [15, 16] has depicted complicated energy flow transactions between two species as energy circuit. A bunch of indicators have been proposed to quantify the properties of this open flow network[17–20], and numeric common properties have been discovered[21–25]. Patten *et al.* proposed a systematic "Ecological Flow Analysis" method to investigate energetic flow networks[26, 27].

Indeed, many basic ideas and approaches of flow analysis on energetic food webs inherit from the economic input-output analysis method[28] which is first proposed by the famous economist Leontief[29, 30]. To quantify the complex economic production processes and the interaction between different economic sectors, an input-output matrix is calculated for an economic system to represent goods flows[31, 32]. Following Leontief's seminal work, Hanon introduced basic notions such as fundamental matrix to ecology for describing the energy flows between species[33]. Therefore, an input-output matrix can also be regarded as an open flow network. Money flow from the final demands compartment, circulate in different sectors of an economic system, and eventually flow to the value added compartment (or goods flow in an inverse direction). Thus, value-added compartment can be regarded as the sink, and final demands can be regarded as the source. The money flow from industry $i$ to industry $j$ is always measured by the uniform currency unit, therefore the total out flow from the source equals the total inflow to the sink, and is identical to the gross domestic output of an economy[31, 32]. Other examples of open flow networks include clickstream networks[34, 35] and trade networks[36]. In summary, open flow network is a very useful tool to depict various open flow systems.

Distance on graph is a very useful concept[37]. Both the shortest path distance[38], resistance distance[39] and

* zhangjiang@bnu.edu.cn; http://www.swarma.org/jake

the mean first-passage distance of a random walker[40–43] can reflect the intrinsic properties of the graph. However, conventional first-passage distance on a graph is based on the basic assumption that the whole network is closed, which means the random walker cannot escape from the network, thus the total number of walkers on the graph is conservative. Nevertheless, the open flow network is an open system. Random walkers can flow into the system from the source and flow out to the sink despite the total number of walkers staying in the network can be also conservative if the flow system is in a steady state. Therefore, the traditional method for closed system cannot be simply extended to open flow networks. It is necessary to extend the distance notions for open flow networks.

This paper is organized as follows. In section II, the flow distance quantities from $i$ to $j$ are defined. The explicit form of each flow distance is expressed in subsection II C. Sub-section II D shows how the distance matrix is calculated on an example flow network. In sub-section III A, we apply our method to energetic food webs, visualize each species by its trophic level, and compare different distances on the food webs. The applications of flow distances on input-output network including network visualization, sector clustering, and vertex centrality are introduced in sub-section III B. Finally, we give a short summary for all the paper and the perspective of flow distances in section IV.

## II. FLOW DISTANCES

In this section, we will present the definitions and calculations of flow distances. Three flow distances, namely first-passage flow distance, total flow distance, and symmetric flow distance, are defined. They all can be expressed by the Markov matrix of the open flow network. To obtain the final expressions, some intermediate concepts including total flow and first-passage flow are needed to be introduced.

### A. Definitions

Consider an open flow network with $N$ common nodes and two special nodes "source" denoted by 0 and "sink" denoted by $N+1$ are added. An $(N+2)\times(N+2)$ matrix $F$ can be used to represent flows, and each entry $f_{ij}$, where $i,j \in 0,1,2,\cdots,N+1$, represents the flow from node $i$ to $j$. Note that the elements in the first column and the last row are all equal 0 because there are no inflow to the source and no out flow for the sink. We also define $f_{i\cdot} = \sum_{j=0}^{N+1} f_{ij}$ as the total out flow from $i$, and $f_{\cdot j} = \sum_{i=0}^{N+1} f_{ij}$ as the total inflow to $j$. In our research, the flow network should be balanced, which means that $f_{\cdot i} = f_{i\cdot}$ for every node $i$ except "source" and "sink". Particularly, we name $f_{i,N+1}$, the flow from $i$ to the sink, as **dissipation**.

Suppose a large number of particles flow along links in the network $F$, the directed flow $f_{ij}$ from $i$ to $j$ is the total number of particles that jump from $i$ directly to $j$ along edge $i \to j$ in each time. The particles may jump from $i$ to $j$ along indirected paths, we define the **first-passage flow** from $i$ to $j$ denoted by $\phi_{ij}$ as the number of particles that reach $j$ in each time step for the first time and have been visited $i$. And the average step that these particles have jumped is defined as the **first-passage flow distance** which is denoted by $l_{ij}$.

Similarly, the **total flow** from $i$ to $j$ denoted as $\rho_{ij}$ is defined as the total number of particles that have been visited $i$ and arrive at $j$ in each time no matter if it is the first time or not. And the average step that these particles have jumped is defined as the **total flow distance** which is denoted by $t_{ij}$.

To understand these quantities better, let's consider the following imaginary experiment. Suppose all the particles passing by node $i$ are dyed red and this color would be washed out once the red particles arrive at node $j$ for the first time. Then the first-passage flow from node $i$ to $j$ is the number of red particles passing by node $j$ in each time. The first-passage flow distance is the average step that these particles have made. Similarly, if the particles passing by $i$ are dyed red but this color would never be washed out, then the number of red particles that pass by $j$ in each time is the total flow, and the average step that these particles have made is the total flow distance.

In this paper, all the matrices are denoted by capital letters, and the their corresponding elements are denoted by lower case of the name of matrices. For example, $F$ denote the flow matrix, and $f_{ij}$ is the element of the $i$th row and $j$th column

### B. Calculation of total flow and first-passage flow

Because the open flow system is in a steady state, and the flow network is balanced, we can define a Markov matrix $M$ as follows,

$$m_{ij} = \frac{f_{ij}}{\sum_{j=1}^{N+1} f_{ij}} \tag{1}$$

and $m_{ij}$ represents the probability of particles jumping from state $i$ to $j$. Note that $\sum_{j=1}^{N+1} m_{ij} = 1$ for any $i$ except $N+1$ because the elements in the last $((N+1)$th) row are all zeros, this is a key difference between open flow network and closed flow network.

According to reference [19], no matter if circulations exist in network, the total flow from $i$ to $j$ can be calculated as:

$$\rho_{ij} = \phi_{0i} u_{ij}, \tag{2}$$

where

$$U = I + M + M^2 + \cdots = (I - M)^{-1} \tag{3}$$

is called fundamental matrix, it is also the inverse of $M$'s laplacian. And $\phi_{0i}$ is the first-passage flow from the source to $i$ which will be calculated in the following paragraphs. $I$ is the identity matrix with size $(N+1)\cdot(N+1)$. We will provide an informal proof for Eq (2).

Equation (2) calculates the total flows along all possible paths from $i$ to $j$. When $i \neq j$, the number of particles that jump from $i$ to $j$ along all possible paths with $k$ steps is $\phi_{0i}(M^k)_{ij}$. Note that particles may flow back to $i$ for several times, $\phi_{0i}$ instead of $f_{\cdot i}$ is adopted because $f_{\cdot i}$ contains the flows back to $i$. If the particle passing by $i$ is dyed, then $\phi_{0i}$ is the number of particles without color marker and will be dyed in each time. Taking summation of $\phi_{0i}(M^k)_{ij}$ from $k=1$ to $\infty$, we can obtain the total flow from $i$ to $j$ along all possible ways. According to the series expansion $MU = M(I-M)^{-1} = M + M^2 + \cdots$ and the identity $(MU)_{ij} = u_{ij}$ when $i \neq j$, Eq. (2) is obtained.

When $i = j$, according to $\rho$'s definition, $\rho_{ii}$ should contain the first-passage flow from the source to $i$, therefore, $\rho_{ii} = \phi_{0i}((MU)_{ii}+1) = \phi_{0i}((MU)_{ii}+I_{ii}) = \phi_{0i}u_{ii}$, then Eq. (2) holds.

Because the total flow from $i$ to $j$ can be divided into two different categories, one is the first-passage flow which contains the particles that arrive at $j$ for the first time, the other is the circulation flow which contains the particles that arrive at $j$ more than once. All the flows are conditioned on starting from $i$. We know that the circulation flow is the summation of flows from $j$ to $j$ along all possible paths, it is calculated as

$$\psi_{ij} = \phi_{ij}(\sum_{k=1}^{\infty} M^k)_{jj} = \phi_{ij}(MU)_{jj}, \tag{4}$$

where $\psi_{ij}$ represents the circulation flow starting from $i$.

Therefore, the total flow from $i$ to $j$ can be expressed as[19]

$$\rho_{ij} = \phi_{ij} + \psi_{ij} = \phi_{ij}u_{jj}. \tag{5}$$

Thus, we obtain the expression for the first-passage flow from $i$ to $j$:

$$\phi_{ij} = \frac{\rho_{ij}}{u_{jj}}. \tag{6}$$

Based on the equations of Eq. (2) and Eq. (6), and note that $\phi_{00} = f_{0\cdot}$ according to the definition, where $f_{0\cdot}$ denotes the total flow from "source" to the whole system, we have

$$\rho_{ij} = \phi_{0i}u_{ij} = \frac{\rho_{0i}}{u_{ii}}u_{ij} = f_{0\cdot}\frac{u_{0i}}{u_{ii}}u_{ij}. \tag{7}$$

And the explicit expression for the first-passage flow is

$$\phi_{ij} = \frac{\rho_{ij}}{u_{jj}} = \phi_{0i}u_{ij}\frac{1}{u_{jj}} = f_{0\cdot}\frac{u_{0i}u_{ij}}{u_{ii}u_{jj}}. \tag{8}$$

### C. Calculation of flow distances

We can deduce the explicit expression of various flow distances once the total flow and first-passage flow expressions are given. First, according to the definition of the total flow from $i$ to $j$ along all possible paths, we have

$$t_{ij} = \sum_{k=1}^{\infty} k p_{ij}^k, \tag{9}$$

where $p_{ij}^k$ denotes the probability that particles transfer from $i$ to $j$ after $k$ steps. One may think $p_{ij}^k = (M^k)_{ij}$, however, it is not true because $p_{ij}^k$ is normalized for all paths with all possible lengths $k$, i.e., $\sum_{k=1}^{\infty} p_{ij}^k = 1$. However, $(M^k)_{ij}$ is normalized for all $j$s, i.e., $\sum_{j=0}^{N+1}(M^k)_{ij} = 1$. We know that the flow from $i$ to $j$ after $k$ steps is $\phi_{0i}(M^k)_{ij}$ and the total flow along all possible paths is $\rho_{ij}$, therefore,

$$p_{ij}^k = \frac{\phi_{0i}(M^k)_{ij}}{\rho_{ij}}. \tag{10}$$

Thus bring this equation to Eq. (9), we have,

$$\begin{aligned} t_{ij} &= \sum_{k=1}^{\infty} k\frac{\phi_{0i}(M^k)_{ij}}{\rho_{ij}} = \frac{\phi_{0i}(\sum_{k=1}^{\infty} kM^k)_{ij}}{\rho_{ij}} \\ &= \frac{\phi_{0i}(MU^2)_{ij}}{\rho_{ij}} = \frac{\phi_{0i}(MU^2)_{ij}}{\phi_{0i}u_{ij}} \\ &= \frac{(MU^2)_{ij}}{u_{ij}}. \end{aligned} \tag{11}$$

In which, we have used the following series expansion:

$$MU^2 = M\left(\frac{1}{I-M}\right)^2 = \sum_{k=1}^{\infty} kM^k. \tag{12}$$

Similarly, we can obtain the expression for first-passage flow distance. First, according to the definition of the first-passage distance from $i$ to $j$, we have

$$l_{ij} = \sum_{k=1}^{\infty} k q_{ij}^k, \tag{13}$$

where $q_{ij}^k$ denotes the probability that particles started from $i$ to $j$ after $k$ steps in the first time. One cannot use $p_{ij}$ (Eq. (10)) because it contains the circulation flow from $j$ to $j$. Let us assume that all the particles arriving at $j$ will be removed from the system, that is to say, we assume that $j$ is another sink, then all the calculations for the total flow distance is correct. To make this point clear, we define a new matrix $M_{-j}$ as:

$$(M_{-j})_{rs} = \begin{cases} m_{rs}, & r \neq j \\ 0, & r = j. \end{cases} \tag{14}$$

And the correct expression for the probability $q_{ij}^k$ is

$$q_{ij}^k = \frac{\phi_{0i}(M_{-j}^k)_{ij}}{\phi_{ij}}. \tag{15}$$

Insert it into Eq. (13), we have

$$l_{ij} = \frac{u_{jj}(M_{-j}U_{-j}^2)_{ij}}{u_{ij}}. \tag{16}$$

According to the Theorem 1 proved in the Supplementary Material, when $u_{ij} \neq 0$ ($i$ connects to $j$), this formula can be reduced to

$$l_{ij} = \frac{(MU^2)_{ij}}{u_{ij}} - \frac{(MU^2)_{jj}}{u_{jj}} = t_{ij} - t_{jj}. \tag{17}$$

Therefore, the difference between $t_{ij}$ and $l_{ij}$ is just the total flow distance from $j$ to $j$. The matrix $(t_{jj})_{N+1,N+1}$ have identical rows. Therefore, we can use a vector $t_j$ to abbreviate the matrix $t_{jj}$, it quantifies the ability of self circulation of each node in the system.

All the flow distances introduced above are asymmetric, however, some real tasks such as nodes clustering, computation of node centrality require symmetric metrics. Commute distance[42, 43] is a classical and famous symmetric distance measure defined by random walk, which is calculated by $l_{ij} + l_{ji}$. However, this definition cannot work when one of $l_{ij}$ or $l_{ji}$ is infinity meaning that $i$ cannot access $j$ or vice versa. Therefore, we define a new symmetric flow distance to avoid this problem:

$$c_{ij} = 2\frac{1}{\frac{1}{l_{ij}} + \frac{1}{l_{ji}}} = \frac{2l_{ij}l_{ji}}{l_{ij} + l_{ji}}. \tag{18}$$

We call $c_{ij}$ symmetric flow distance, it is a mixing of $l_{ij}$ and $l_{ji}$. Suppose $l_{ij} = \infty$, then $c_{ij} = 2l_{ji}$ which is well-defined. When $l_{ij} = l_{ji}$, $c_{ij} = l_{ij} = l_{ji}$. Therefore, $c_{ij}$ is a reasonable symmetric distance.

### D. Calculation on an example network

Before applying our method to real open flow networks, we would like to present the computations of flow distances on a small example network and compare with other distances on graph. The example network is shown in Figure 1. There are 7 nodes including the source and the sink. All the flows are denoted on the edges. We present the first-passage flow distances matrix $L$ in the following equation

$$\begin{bmatrix} 0 & 1 & 2.15 & 2.27 & 3.20 & 3.40 & 3.94 \\ \infty & 0 & 1.15 & 1.27 & 2.20 & 2.40 & 2.94 \\ \infty & \infty & 0 & 2.25 & 1.05 & 1.25 & 2.13 \\ \infty & \infty & 1 & 0 & 2.05 & 2.25 & 1.61 \\ \infty & \infty & 3 & 2 & 0 & 1 & 1.60 \\ \infty & \infty & 2 & 1 & 3.05 & 0 & 1.20 \\ \infty & \infty & \infty & \infty & \infty & \infty & 0 \end{bmatrix} \tag{19}$$

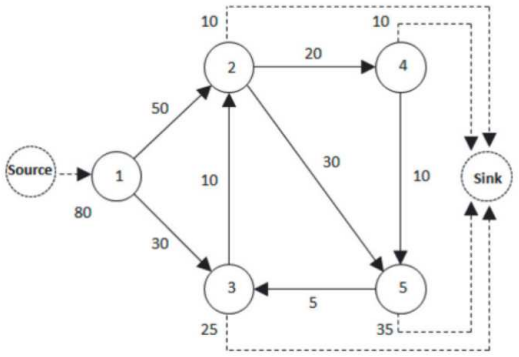


FIG. 1. An example open flow network

and the total flow distances matrix $T$

$$\begin{bmatrix} 0 & 1 & 2.23 & 2.35 & 3.23 & 3.48 & 3.94 \\ \infty & 0 & 1.23 & 1.35 & 2.23 & 2.48 & 2.94 \\ \infty & \infty & 0.08 & 2.33 & 1.08 & 1.33 & 2.13 \\ \infty & \infty & 1.08 & 0.08 & 2.08 & 2.33 & 1.61 \\ \infty & \infty & 3.08 & 2.08 & 0.03 & 1.08 & 1.60 \\ \infty & \infty & 2.08 & 1.08 & 3.08 & 0.08 & 1.20 \\ \infty & \infty & \infty & \infty & \infty & \infty & 0 \end{bmatrix} \tag{20}$$

Note that there are many $\infty$ entries in both $L$ and $T$ because the corresponding node pairs have no connected path. Another interesting phenomenon is all the elements in $T$ are larger than the corresponding entries in $L$. And the difference $(T - L)$ is:

$$\begin{bmatrix} 0 & 0 & 0.08 & 0.08 & 0.03 & 0.08 & 0 \\ & 0 & 0.08 & 0.08 & 0.03 & 0.08 & 0 \\ & & 0.08 & 0.08 & 0.03 & 0.08 & 0 \\ & & 0.08 & 0.08 & 0.03 & 0.08 & 0 \\ & & 0.08 & 0.08 & 0.03 & 0.08 & 0 \\ & & 0.08 & 0.08 & 0.03 & 0.08 & 0 \\ & & & & & & 0 \end{bmatrix} \tag{21}$$

The empty entries have no numeric value because $\infty - \infty$ is indefinite. All the elements in the same column are identical which are the average flow distances from $i$ to $i$ for $i = 2, 3, 4, 5$. And because $2, 3, 5$ are in the same cycle $2 \to 5 \to 3$ and $2 \to 4 \to 5 \to 3$, they have the same values of $t_{jj}$.

Next, we compare our first-passage flow distance $l_{ij}$ with shortest path distance and first-passage distance based on random walks[41] on the closed version of the same network. In the latter comparison, "source" and "sink" are excluded so that the network is closed. For the random walkers in the closed network, the transition probability between $i$ and $j$ is the fraction between $f_{ij}$ and $(f_{i\cdot} - f_{i,N+1})$, the total out flows from $i$ excluding the dissipation from $i$ . For example, the transition probability from 2 to 4 is $20/(20 + 30) = 0.4$ but not $20/(20 + 30 + 10) = 0.33$. The results are shown in Table I.

TABLE I. Comparisons among three kinds of distances on selected node pairs

| | 1→3 | 2→3 | 1→4 | 2→4 |
|---|---|---|---|---|
| Shortest path | 1 | 2 | 2 | 1 |
| Closed FPD | 2.5 | 2.4 | 6.875 | 5.5 |
| Open FPD | 1.274 | 2.25 | 2.2 | 1.055 |

As we expected, shortest path lengths are much shorter than the other two distances because they only consider the shortest paths, as a result, this distance always under-estimates the distances of random particles in flows. Comparing two first-passage distances is more interesting. Closed First Passage Distance (Closed FPD) is always larger than Open First Passage Distance (Open FPD) because dissipations are not considered in Closed FPD. For example, the Closed FPD from 1 to 4 is larger than the Open FPD in almost 3 times because the longer circulated path $2 \to 5 \to 3 \to 2$ has much higher probability when the dissipation from node 2 is neglected (0.6 rather than 0.5).

## III. EMPIRICAL STUDIES

In this section, we will apply our flow distances on two kinds of networks: 18 energetic food webs (energy flow information between species is included) and the input-output network of U.S. The raw data of food webs is from the following open data source[44]. And the input-output network data is from[45].

### A. Food web

Trophic level is an important concept in food webs, it characterizes one species' distance from the source (sun light) along the food chain. However, when the food web is an entangled network, calculating the shortest path from the source always under-estimates the trophic level of a given species because non-shortest paths may have much longer distances from the source. Therefore, we quantify trophic levels of different species by the concept of first-passage distance from the source[20], that is $l_{0i}$ for any species $i$. This distance is reasonable because it contains the information of weights and all possible energetic path ways from the source. Figure 2.a visualizes the trophic levels of 125 biological species in Baydry food web. Producer species locate at the area close to the source, and higher level consumers locate in the peripheries.

Next, we calculate several distances in the level of the entire network. The first distance is the first-passage distance from the source to the sink ($l_{0,N+1}$). This distance quantifies the average number of steps of a random particle in its all life span. The second distance is the mean value of the elements in matrix $L$ except for the infinite elements. We calculate these distances for all the collected energetic food webs, and to observe how the distances change with network size.

Figure 3 shows various distances change with number of edges of networks. We find that the average value of $l_{ij}$ has similar trend with the average path length $l_{0,N+1}$ from the source to the sink. Shortest path length is always shorter than the average $l$ and $l_{0,N+1}$ because it does not consider the average behaviour of random walkers. There is a slightly trend that the network lengths increase with network size.

### B. Input-output network

Input-output network is another kind of flow network. Each industrial sector corresponds to a vertex, and an input from one sector to another can be considered as a flow. However, there are two kinds of views to represent an input-output network as a flow network. If we consider material flow, then the input from sector $i$ to $j$ should be understood as a flow from $i$ to $j$. However the flow may be from $j$ to $i$ if money flow is considered. We adopt the viewpoint of money flow in this paper because the flow of money in different sectors resembles random walkers in open flow networks. In this way, the final demand sector is the source of money flows, and the value added sector is the sink. We choose the input-output data from United States in 2000 as an example to calculate various flow distances.

First, it is curious to calculate the economic "trophic levels" ($l_{0,i}$) of different sectors (see Figure 2.b). The sectors with shorter distances from the center are closer to the source, therefore they are more easily to be affected by the final demand. Any fluctuations of demands or price can be transferred to the sectors with lower "trophic levels".

Second, we can use flow distances to calculate similarity between different sectors. Because it is much easier to deal with symmetric similarity, we use the distances $c_{ij}$ instead of $l_{ij}$ here. With this symmetric measure, we can cluster sectors by using the standard hierarchical clustering techniques[46]. The result is visualized by Figure 4. In this figure, similar or related sectors are gathered closely, like *Public admin* and *Health & socialwork*, *Ming* and *Fuel*. We also find that *Real estate* sector is close to *Finance* sector, which means real estate has tight relation with finance in U.S. The clustering results have good agreement with our common sense of industrial sectors.

Furthermore, the symmetric measure $c_{ij}$ can be used to measure the centrality of each node because if $i$'s average $c_{ij}$ for different $j$ is shorter then $i$ must have tight connections with all other nodes. Formally, we define the centrality of node $i$ as

$$\bar{c_i} = \frac{\sum_{j=1}^{N} c_{ij}}{N} \tag{22}$$

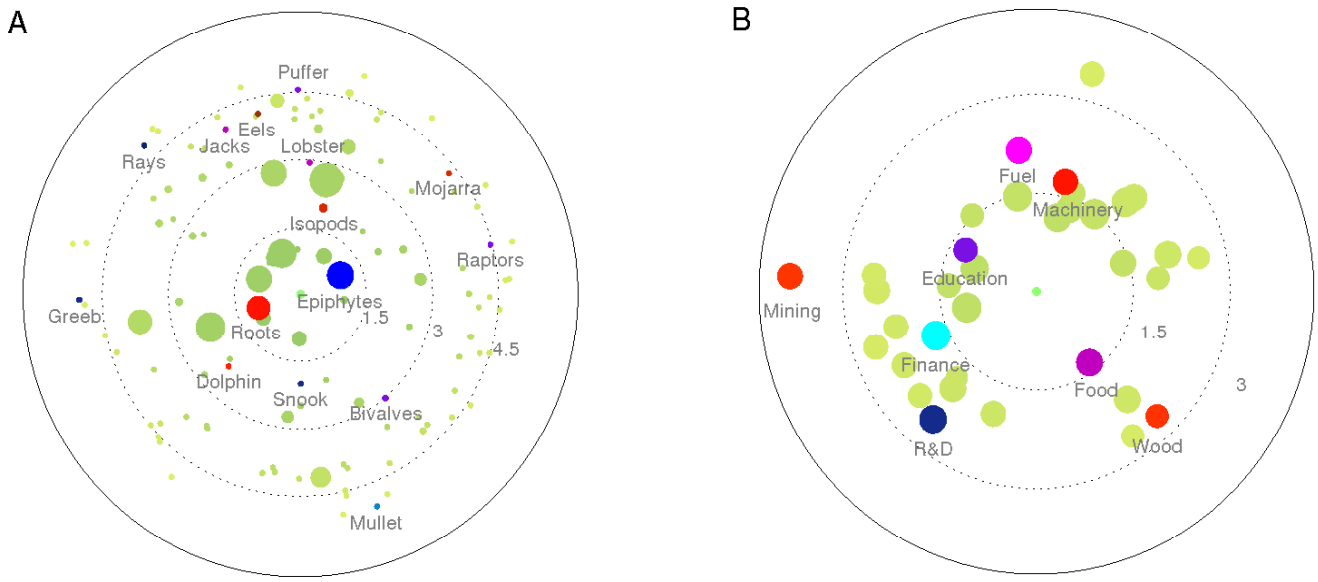


FIG. 2. Trophic levels of species in Baydry community (a) and industrial sectors of U.S. input-output network in 2000 (b). The polar radii, i.e., the distances between every node and the center are proportional to nodes' trophic levels, the polar angles are randomly assigned, and the sizes of nodes are proportional to the logarithmic volumes of the total throughflow for each node ($f_{i\cdot}$). The colors are assigned randomly.

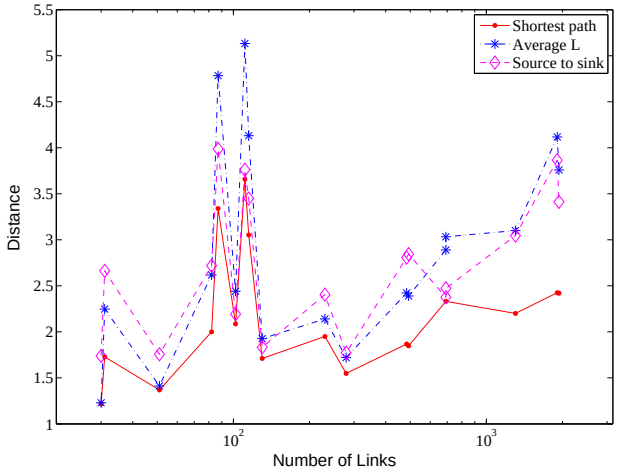


FIG. 3. Three kinds of distances for all collected energetic food webs. All food webs are sorted according to their number of edges in an increasing order

Thus, the shorter is $i$'s $\bar{c_{ij}}$, the more central position it has in the whole economic system. We color different nodes in Fig. 4 by $\bar{c_i}$. The color depth increases as $c_i$ decreases. We find that *Trade* and *Public admin.* sectors are more central than other sectors in U.S., and *Agriculture* and *Ming* sectors are less important than the average.

Finally, we calculate the vector $t_i$ for all $i$. It is defined as the average steps of a random walker who starts from $i$ and finally returns to $i$ again. This measure indicates the re-cycle capability of a sector in the sense of money flow. Therefore, less $t_i$ implies larger capability of self-maintenance of this sector. In Table II, we show the top 5 and bottom 5 sectors in the decreasing order of $t_i$ in the United States.

TABLE II. List of sectors sorted by $t_i$

| Rank | Sectors in USA |
|---|---|
| 1 | Motor vehicles, trailers and semi-trailers |
| 2 | Finance and insurance |
| 3 | Basic metals |
| 4 | Chemicals and chemical products |
| 5 | Agriculture, hunting, forestry and fishing |
| ⋮ | |
| 32 | Electricity, gas and water supply |
| 33 | Hotels and restaurants |
| 34 | Construction |
| 35 | Education |
| 36 | Health and social work |

The top five sectors are more likely connected to other sectors in the economy. Through analysing the flux matrix $F$, we find that they have less fractions of flows from the source or to the sink. On the contrary, the last five

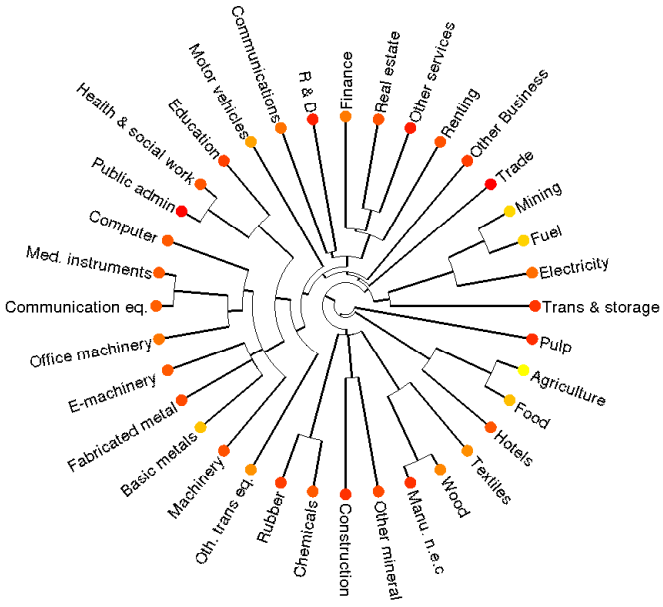


FIG. 4. Hierarchical clustering of different industrial sections in U.S. Colors represent node centrality. All the sector names are abbreviated, and full names can be referred to the Supplementary Material.

sectors are all major in providing services or products for final demand.

## IV. DISCUSSION

In this paper, we introduce flow distances in various open flow networks. These distances characterize interactions between different nodes, and all the distances can be expressed explicitly by the Markov matrix. We give several examples on potential applications of flow distances on energetic food webs and input-output network. Trophic level as an important conception introduced in food web ecology should be applied to other open flow networks. Usually, the nodes with lower trophic levels are more probable to be influenced by the source easily. Second, we can use flow distances to cluster nodes because the symmetric distance $c_{ij}$ can be regarded as a kind of similarity measure. We also use $c_{ij}$ to compare node centrality between different nodes. Vector $t_i$ can be used as an indicator to compare the in-dependency of different node. Because all the flow distances reflect the nature of random walk in an open flow network, they combine the topology and flow dynamics on the network together. Therefore, these distances must have very wide application background.

Certainly, the applications of flow distances should not be limited by the examples listed in this paper. First, open flow networks are combinations of network structure and random walk dynamics, thus visualizing these networks needs particular method. Besides placing different nodes on a space by their "trophical levels" directly, we can embed the flow network into a Euclidean space according to distance $c_{ij}$ such that the Euclidean distance of any given pair $i$ and $j$ is as close as their $c_{ij}$. This embedding problem can be solved by optimizing the places of each node in the Euclidean space. And the patterns of the nodes distributed in the space may help us to understand the flow network structure in an intuitive way. However, how to visualize the open flow networks to reflect the characteristics of the directionality and weights of edges is another important issue deserving for further studies. Second, open flow networks always resemble tree structures that are hierarchical and possessing multi-level structures. How to partition a flow network into several smaller sub-structures, and how to coarse-grain these structures is also an interesting problem. It is reasonable to develop a novel method based on flow distances discussed in this paper to partition and coarse-grain. Third, the flow distances metrics and network embedding can help us to understand some underlying dynamical processes on the network in a geometric way[37].

Flow distances can obviously applied to other open flow networks, and may facilitate us to compare them. Trade flow network, traffic flow network, attention flow networks are all very important examples. Application of flow distances on these networks may reveal important common patterns.

The current flow distances metrics also have shortcomings. The computational complexity will increase fast as the size of the network because the matrices $U$ and $L$ are non-sparse when the network is large. Therefore, the approximate algorithm of flow distances is very necessary and urgent. Additionally, all the flow distances metrics are average values of various paths of particles, the variances of these paths cannot be reflected on these metrics. New indicators are needed to represent the fluctuations of different paths. All these problems deserve further studies.

[1] M. Newman, A. L. Barabási, and D. J. Watts, *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, 2006).
[2] A. L. Barabási and E. Bonabeau, Sci. Am., 50 (2003).
[3] S. H. Strogatz, Nature, **410**, 268 (2001), PMID: 11258382.
[4] R. Albert and A. L. Barabási, Rev. Mod. Phys., **74**, 47 (2002).
[5] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, Proc. Natl. Acad. Sci. U. S. A., **101**, 3747C (2004).
[6] S. Horvath, *Weighted Network Analysis. Applications in*

*Genomics and Systems Biology* (Springer, 2011).
[7] J. Bang-Jensen and G. Gutin, *Digraphs: Theory, Algorithms and Applications* (Springer, 2000) ISBN 1-85233-268-9.
[8] A. S. Asratian, T. M. J. Denley, and R. Haggkvist, *Bipartite Graphs and their Applications* (Cambridge University Press, 1998) cambridge Tracts in Mathematics 131.
[9] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, Nature, **464**, 1025 (2010).
[10] K.-M. Lee, J. Y. Kim, W.-K. Cho, K.-I. Goh, and I.-M. Kim, New. J. Phys., **14** (2012), doi:10.1088/1367-2630/14/3/033027.
[11] P. Holme and J. Saramäki, Phys. Rep., **519**, 97 (2012).
[12] G. Nicolis and I. Prigogine, *Self-organization in nonequilibrium systems* (Wiley, New York, 1977).
[13] G. West and J. Brown, J. Exp. Biol., **208**, 1575 (2005).
[14] J. Banavar, A. Maritan, and A. Rinaldo, Nature, **399**, 130 (1999).
[15] H. Odum, *System Ecology; an introduction* (John Wiley & Sons Inc, 1983).
[16] H. Odum, Science, **242**, 1132 (1988).
[17] B. D. Fath and B. C. Patten, Ecosystems, **2**, 167 (1999).
[18] J. T. Finn, J. Theor. Biol., **56**, 363 (1976).
[19] M. Higashi, B. C. Patten, and T. P. Burns, Ecol. Model., **66**, 1 (1993).
[20] S. Levine, J. Theor. Biol., **83**, 195 (1980), ISSN 0022-5193.
[21] R. E. Ulanowicz, *Ecology, the Ascendent Perspective* (Columbia University Press, New York, 1997).
[22] R. E. Ulanowicz, Comput. Biol. Chem., **28**, 321 (2004).
[23] J. Zhang and L. Guo, J. Theor. Bio., **264**, 760 (2010).
[24] J. Zhang and L. Wu, PLoS ONE, **8**, e72525 (2013).
[25] J. Zhang and Y. Feng, Physica A, **405**, 278 (2014).
[26] B. C. Patten, Am. Nat., **119**, 179 (1982).
[27] B. C. Patten, Am. Zool., **21**, 845 (1981).
[28] M. Higashi, Ecol. Model., **32**, 137 (1986).
[29] W. W. Leontief, *Structure of American economy,1919-1929:An Empirical Application of Equilibrium Analysis* (Harvard University Press, 1941).
[30] W. W. Leontief, *Input-output economics* (Oxford University Press, 1986).
[31] T. Ten Raa, *The economics of input-output analysis* (Cambridge University Press, Cambridge, 2005) ISBN 0521841798 9780521841795 052160267X 9780521602679.
[32] R. E. Miller and P. D. Blair, *Input-output analysis: foundations and extensions* (Cambridge University Press, Cambridge [England]; New York, 2009) ISBN 9780521517133 0521517133 0521739020 9780521739023.
[33] B. Hannon, J. Theor. Biol., **41**, 535 (1973).
[34] L. Wu, J. Zhang, and M. Zhao, PLoS ONE, **9**, e102646 (2013).
[35] L. Wu and J. Zhang, Eur. Phys. J. B, **86**, 266 (2013).
[36] P. Shi, J. Zhang, and J. Luo, PLoS ONE, **9**, e98247 (2014).
[37] D. Brockmann and D. Helbing, Science (2013).
[38] B. V. Cherkassky, A. V. Goldberg, and T. Radzik, Math. Program., **73**, 129 (1996).
[39] D. J. Klein and M. J. Randić, J. Math. Chem., **12**, 81 (1993).
[40] J. D. Noh and H. Rieger, Phys. Rev. Lett., **92**, 118701 (2004).
[41] F. Blöchl, F. J. Theis, F. Vega-Redondo, and E. ON. Fisher, Phys. Rev. E., **83**, 046127 (2011).
[42] L. Lovász, Bolyai. Soc. Math. Stud., **2**, 1 (1993).
[43] P. Tetali, J. Theoret. Probab. (1991).
[44] `http://vlado.fmf.uni-lj.si/pub/networks/data/bio/foodweb/fo` (2006).
[45] `http://www.oecd-ilibrary.org/industry-and-services/data/sta`
[46] S. Johnson, Psychometrika, **32**, 241 (1967).

# Supplementary Material for Flow Distances on Open Flow Networks

Liangzhu Guo,[1] Xiaodan Lou,[1] Peiteng Shi,[2] Jun Wang,[2] Xiaohan Huang,[2] and Jiang Zhang[1,*]

[1] *School of Systems Sciences, Beijing Normal University, Beijing, China*
[2] *Science and Technology on Information Systems Engineering Laboratory, National University of Defence Technology, Changsha, China*

(Dated: January 27, 2015)

## I. PROOF OF A THEOREM

In this appendix, we will prove Eq. (17). But before that, several lemmas are needed to be proved at first.

**Lemma 1**: The following equation is true:

$$I = (I - M)U. \tag{1}$$

**Proof**: It is obvious according to the definition of $U = (I - M)^{-1}$.

**Lemma 2**: The following equation is true:

$$(I - M)U = (I - M_{-d})U_{-d} = I. \tag{2}$$

Where, $M_{-d}$ is the matrix when the $d$th row of matrix $M$ is set to zero. Thus,

$$M = M_{-d} + \Delta M, \tag{3}$$

where

$$(\Delta M)_{ij} = \begin{cases} m_{ij}, & i = d \\ 0, & i \neq d \end{cases} \tag{4}$$

Correspondingly, $U_{-d}$ is

$$U_{-d} = I + M_{-d} + M_{-d}^2 + \cdots = (I - M_{-d})^{-1}. \tag{5}$$

**Proof**: This is also obvious according to Lemma 1.

**Lemma 3**: The following equation holds for any $d, i, j$ belongs to $[1, N]$:

$$(U_{-d})_{ij} = u_{ij} - (U_{-d})_{id}u_{dj} \tag{6}$$

$$= u_{ij} - \frac{u_{id}}{u_{dd}}u_{dj} - \frac{u_{id}}{u_{dd}}\delta_{dj}. \tag{7}$$

Where,

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{8}$$

**Proof**: According to Lemma 2, we have

$$U - MU = U - (M_{-d} + \Delta M)U = U_{-d} - M_{-d}U_{-d} \tag{9}$$

$$\Rightarrow U - U_{-d} = (M_{-d} + \Delta M)U - M_{-d}U_{-d} = M_{-d}(U - U_{-d}) + \Delta MU$$

$$\Rightarrow (I - M_{-d})(U - U_{-d}) = \Delta MU \tag{10}$$

$$\because\ U_{-d}(I - M_{-d}) = I \tag{11}$$

$$\therefore\ U - U_{-d} = U_{-d} \cdot \Delta M \cdot U \tag{12}$$

According to the definition of $\Delta M$, and according to the fact that

$$\sum_k m_{ik}u_{kj} = u_{ij} - \delta_{ij}, \tag{13}$$

we can expand $U_{-d} \cdot \Delta M \cdot U$ as

$$U_{-d} \cdot (\Delta M \cdot U)_{ij} = (U_{-d})_{id} \cdot \sum_k m_{dk}u_{kj} = (U_{-d})_{id}(u_{dj} - \delta_{dj}). \tag{14}$$

So, we get

$$u_{ij} - (U_{-d})_{ij} = (U_{-d})_{id}(u_{dj} - \delta_{dj}) \tag{15}$$

In the above equation, if we let $j = d$, then

$$u_{id} - (U_{-d})_{id} = (U_{-d})_{id}(u_{dd} - 1) \tag{16}$$

Thus,

$$(U_{-d})_{id} = \frac{u_{id}}{u_{dd}}. \tag{17}$$

Insert it into Equation (15), we have

$$u_{ij} - (U_{-d})_{ij} = \frac{u_{id}}{u_{dd}}(u_{dj} - \delta_{dj}). \tag{18}$$

At last, rearrange this equation, we obtain

$$(U_{-d})_{ij} = u_{ij} - \frac{u_{id}}{u_{dd}}u_{dj} - \frac{u_{id}}{u_{dd}}\delta_{dj}. \tag{19}$$

**Lemma 4**: Based on these lemmas, we can get such equation:

$$(U_{-j}^2)_{ij} = \frac{(U^2)_{ij}}{u_{jj}} - \frac{u_{ij}}{u_{jj}^2}(U^2)_{jj} + \frac{u_{ij}}{u_{jj}}. \tag{20}$$

* zhangjiang@bnu.edu.cn; http://www.swarma.org/jake

**Proof**: Expand $U^2$ into elements, and substitute Lemma 3 into it, then

$$\begin{aligned}(U^2_{-j}) &= \sum_k (U_{-j})_{ik}(U_{-j})_{kj} \\ &= \sum_k (u_{ik} - \frac{u_{ij}}{u_{jj}}u_{jk} + \frac{u_{ij}}{u_{jj}}\delta_{jk})\frac{u_{kj}}{u_{jj}} \\ &= \sum_k \frac{u_{ik}u_{kj}}{u_{jj}} - \sum_k \frac{u_{ij}}{u^2_{jj}}u_{jk}u_{kj} + \frac{u_{ij}}{u_{jj}} \\ &= \frac{(U^2)_{ij}}{u_{jj}} - \frac{u_{ij}}{u^2_{jj}}(U^2)_{jj} + \frac{u_{ij}}{u_{jj}}. \end{aligned} \tag{21}$$

**Theorem 1**: Equation (17) in the main text or the following equation

$$\frac{1}{u_{ij}}[(MU^2)_{ij} - u_{jj}(M_{-j}U^2_{-j})_{ij}] = \frac{(MU^2)_{jj}}{u_{jj}} \tag{22}$$

holds when $u_{ij} \neq 0$.

**Proof**: Substitute $M \cdot U^2 = U^2 - U$ and $M_{-j} \cdot U^2_{-j} = U^2_{-j} - U_{-j}$ into Lemma 1 and Lemma 3, it can be proved.

$$\begin{aligned}&\frac{1}{u_{ij}}[(MU^2)_{ij} - u_{jj}(M_{-j}U^2_{-j})_{ij}] \\ &= \frac{1}{u_{ij}}\big[(U^2 - U)_{ij} - u_{ij}(U^2_{-j} - U_{-j})_{ij}\big] \\ &= \frac{1}{u_{ij}}\left\{(U^2)_{ij} - u_{ij} - u_{jj}\left[\frac{(U^2)_{ij}}{u_{jj}} - \frac{u_{ij}}{u^2_{jj}}(U^2)_{jj}\right]\right)\right\} \\ &= \frac{1}{u_{ij}}\big[(U^2)_{ij} - u_{ij} - (U^2)_{ij} + \frac{u_{ij}}{u_{jj}}(U^2)_{jj}\big] \\ &= \frac{(U^2)_{jj} - u_{jj}}{u_{jj}} \\ &= \frac{(MU^2)_{jj}}{u_{jj}} \end{aligned} \tag{23}$$

## II. NAME LIST FOR SECTORS OF INPUT-OUTPUT NETWORK

Sector names in Figure 4 are abbreviated. The full names corresponded are depict in Table I.

TABLE I. Sector full names

| Abbreviations | Full names |
|---|---|
| Agriculture | Agriculture, hunting, forestry and fishing |
| Mining | Mining and quarrying |
| Food | Food products, beverages and tobacco |
| Textiles | Textiles, textile products, leather and footwear |
| Wood | Wood and products of wood and cork |
| Pulp | Pulp, paper, paper products, printing and publishing |
| Fuel | Coke, refined petroleum products and nuclear fuel |
| Chemicals | Chemicals and chemical products |
| Rubber | Rubber and plastics products |
| Other mineral | Other non-metallic mineral products |
| Basic metals | Basic metals |
| Fabricated metal | Fabricated metal products except machinery and equipment |
| Machinery | Machinery and equipment n.e.c |
| Office machinery | Office, accounting and computing machinery |
| E-machinety | Electrical machinery and apparatus n.e.c |
| Communication eq. | Radio, television and communication equipment |
| Med. instruments | Medical, precision and optical instruments |
| Motor vehicles | Motor vehicles, trailers and semi-trailers |
| Oth. trans eq. | Other transport equipment |
| Manu. n.e.c | Manufacturing n.e.c; recycling |
| Electricity | Electricity, gas and water supply |
| Construction | Construction |
| Trade | Wholesale and retail trade; repairs |
| Hotels | Hotels and restaurants |
| Trans & storage | Transport and storage |
| Communications | Post and telecommunications |
| Finance | Finance and insurance |
| Real estate | Real estate activities |
| Renting | Renting of machinery and equipment |
| Computer | Computer and related activities |
| R&D | Research and development |
| Other Business | Other Business Activities |
| Public admin | Public admin. and defence; compulsory social security |
| Education | Education |
| Health & social work | Health & social work |
| Other services | Other community, social and personal services |
| Private households | Private households with employed persons |