Supplementary Material for Triad-based Comparison and Signatures of Directed Networks

Xiaochuan Xu and Gesine Reinert

October 11, 2018

1 Calculation of the optimal c in $NetEMD$

In this section we prove that for calculating $NetEMD$, the optimal $c \in \mathbb{R}$ that minimises the shifted EMD can only take values in the set of discontinuities of the empirical cumulative distribution function;

Suppose $F(x)$ is an empirical cumulative distribution function based on $n+1$ distinct observations $\alpha_0, \alpha_1, \cdots, \alpha_n$ where α_i has multiplicity f_i, and $G(x)$ is an empirical cumulative distribution function based on $m+1$ distinct observationsand $\beta_0, \beta_1, \cdots, \beta_m$ where β_j has multiplicity g_j. Setting $\alpha_{n+1} = \infty$ and $\beta_{m+1} = \infty$,

$$F(x) = \sum_{i=0}^{n} f_i \cdot \mathbb{1}[\alpha_i \leq x < \alpha_{i+1}] \quad \text{and} \quad G(x) = \sum_{j=0}^{m} g_j \cdot \mathbb{1}[\beta_j \leq x < \beta_{j+1}].$$

Letting $g_j^{(c)} = g_j$ and $\beta_j^{(c)} = \beta_j - c$,

$$G^{(c)}(x) = G(x+c) = \sum_{j=0}^{m} g_j \cdot \mathbb{1}[\beta_j \leq x+c < \beta_{j+1}] = \sum_{j=0}^{m} g_j^{(c)} \cdot \mathbb{1}[\beta_j^{(c)} \leq x < \beta_{j+1}^{(c)}].$$

Now refine the partitions $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_{n+1})$ and $\beta^{(c)} = (\beta_0^{(c)}, \beta_1^{(c)}, \ldots, \beta_{m+1}^{(c)})$ to give a joint partition $\gamma = (\gamma_0, \ldots, \gamma_K)$ so that both α and $\beta^{(c)}$ are coarsenings of γ. Then we can write

$$|F(x) - G^c(x)| = \sum_{k=0}^{K} h_k \cdot \mathbb{1}[\gamma_{k+1} - \gamma_k] \qquad (*)$$

with $h_k = h_k(\alpha, \beta^{(c)}) \in \mathbb{N}_0$. From (*),

$$\int_{-\infty}^{+\infty} |F(x) - G^{(c)}(x)| dx = H(c)$$

is a piece-wise linear function of c and $lim_{c\to-\infty}H(c) = lim_{c\to\infty}H(c) = \infty$. Therefore, there exists c^* that minimizes $H(c)$ and $c^* \in [min\{\beta_m - \alpha_0, \beta_0 - \alpha_n\}, max\{\beta_m - \alpha_0, \beta_0 - \alpha_n\}]$.

To show that c takes on one of the discontinuity points of F or G we proceed by contradiction. Assume that for a fixed c_0 such that $\forall j \in \{0, 1, \ldots, m\}$ there is an index $i(j)$ with $\alpha_{i(j)} < \beta_j^{(c_0)} < \alpha_{i(j)+1}$. Then there exists Δc, s.t. for all j it holds that $\alpha_{i(j)} < \beta_j^{(c_0+\Delta_j c)} < \alpha_{i(j)+1}$. Therefore, $h_k(\alpha, \beta^{(c_0)}) = h_k(\alpha, \beta^{(c_0+\Delta c)})$. We have

$$\begin{aligned}\int_{-\infty}^{+\infty} |F - G^{(c_0+\Delta c)}| dx &= \sum_{k=0}^{K} h_k(\alpha, \beta^{(c_0+\Delta c)})[\gamma_{k+1}(\alpha, \beta^{(c_0+\Delta c)}) - \gamma_k(\alpha, \beta^{(c_0+\Delta c)})] \\ &= \sum_{k=0}^{K} h_k(\alpha, \beta^{(c_0)})[\gamma_{k+1}(\alpha, \beta^{(c_0+\Delta c)}) - \gamma_k(\alpha, \beta^{(c_0+\Delta c)})]\end{aligned}$$

where

$$\begin{aligned}\gamma_k(\alpha, \beta^{c_0+\Delta c}) &= \begin{cases} \alpha_i, & \text{if} \quad \gamma_k(\alpha, \beta^{c_0}) = \alpha_i \\ \beta_j^{c_0+\Delta c}, & \text{if} \quad \gamma_k(\alpha, \beta^{c_0}) = \beta_j^{c_0} \end{cases} \\ &= \begin{cases} \alpha_i, & \text{if} \quad \gamma_k(\alpha, \beta^{c_0}) = \alpha_i \\ \beta_j^{c_0} - \Delta c, & \text{if} \quad \gamma_k(\alpha, \beta^{c_0}) = \beta_j^{c_0}. \end{cases}\end{aligned}$$

Therefore, $\int_{-\infty}^{+\infty} |F - G^{c_0+\Delta c}| dx$ depends linearly on Δc, so that

$$\int_{-\infty}^{+\infty} |F(x) - G^{c_0+\Delta c}(x)| dx = \int_{-\infty}^{+\infty} |F(x) - G^{c_0}(x)| dx + B \cdot \Delta c$$

Hence, we can find Δc (positive or negative), s.t.

$$\int_{-\infty}^{+\infty} |F(x) - G^{c_0+\Delta c}(x)| dx < \int_{-\infty}^{+\infty} |F(x) - G^{c_0}(x)| dx$$

which indicates that c_0 is not optimal. In conclusion, if c^* is optimal, then $\alpha_i = \beta_j^{c^*}$ for at least one pair of i, j.

Example. Here is an example; p is a distribution on 2 points, and q is a distribution on 3 points. Here $F_{\bar{p}}$ and $G_{\bar{q}}$ are the normalized cumulative density function

of $p(x)$ and $q(x)$ respectively.

$$p(x) = \begin{cases} \frac{1}{3}, & x = 0 \\ \frac{2}{3}, & x = 1 \end{cases}, \qquad q(x) = \begin{cases} \frac{1}{9}, & x = -1 \\ \frac{7}{9}, & x = 0 \\ \frac{1}{9}, & x = 1 \end{cases}$$

$$F_{\bar{p}} = \begin{cases} 0, & x < 0 \\ \frac{1}{3}, & 0 \le x < \frac{3\sqrt{2}}{2} \\ 1, & x \ge \frac{3\sqrt{2}}{2} \end{cases}, \qquad G_{\bar{q}} = \begin{cases} 0, & x < -\frac{3\sqrt{2}}{2} \\ \frac{1}{9}, & -\frac{3\sqrt{2}}{2} \le x < 0 \\ \frac{8}{9}, & 0 \le x < \frac{3\sqrt{2}}{2} \\ 1, & x \ge \frac{3\sqrt{2}}{2} \end{cases}$$

The discontinuity points of $F_{\bar{p}}$ and $G_{\bar{q}}$ are $\{0, \frac{3\sqrt{2}}{2}\}$ and $\{-\frac{3\sqrt{2}}{2}, 0, \frac{3\sqrt{2}}{2}\}$ respectively. Therefore, the optimal c can only take value in $\{-\frac{3\sqrt{2}}{2}, 0, \frac{3\sqrt{2}}{2}, 3\sqrt{2}\}$ and we obtain the *NetEMD* distance, which is $\frac{\sqrt{2}}{2}$, through direct evaluation for these four values.

2 An example calculation for the triad degree distribution

Figure 1 shows an example network; the Triad degree distribution is given in Table 1, and the cumulative triad degree distribution of orbit 14, abbreviated by CDF_{14}, is shows in Figure 2. Each orbit i results in a distribution, so in total there are 30 distributions for each network.

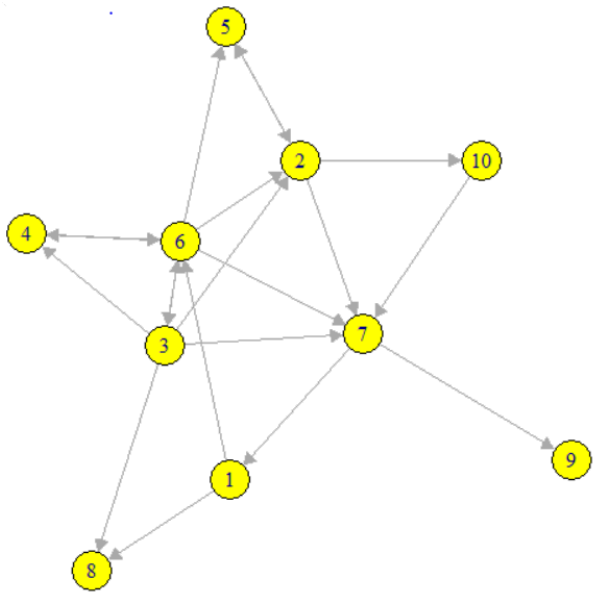

Figure 1: Example: Nodes 1 to 10 are involved in 2, 2, 3, 0, 0, 2, 1, 0, 0, 2 triads of orbit 14 respectively.

k_i	0	1	2	3
$P(k_i)$	0.4	0.1	0.4	0.1

Table 1: Triad degree distribution of orbit 14

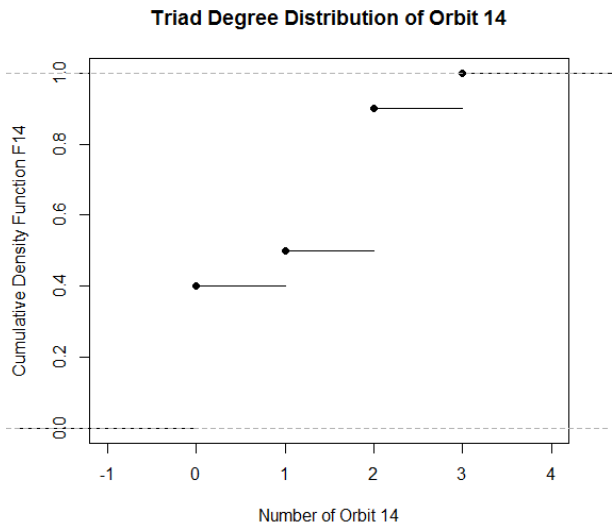

Figure 2: Cumulative distribution function of the triad degree distribution of orbit 14

3 Detailed results for Case Study I

3.1 Cluster dendograms for synthetic data

The cluster dendrogram of $TriadEuclid$ is shown in Figure 3 and that of $TriadEMD$ is shown in Figure 4. We cluster them into 3 groups and each red box represents one group. In both methods, all the networks are classfied correctly, so all the Adjusted Rand Indices and the Nearest Neighbour Scores are 1.

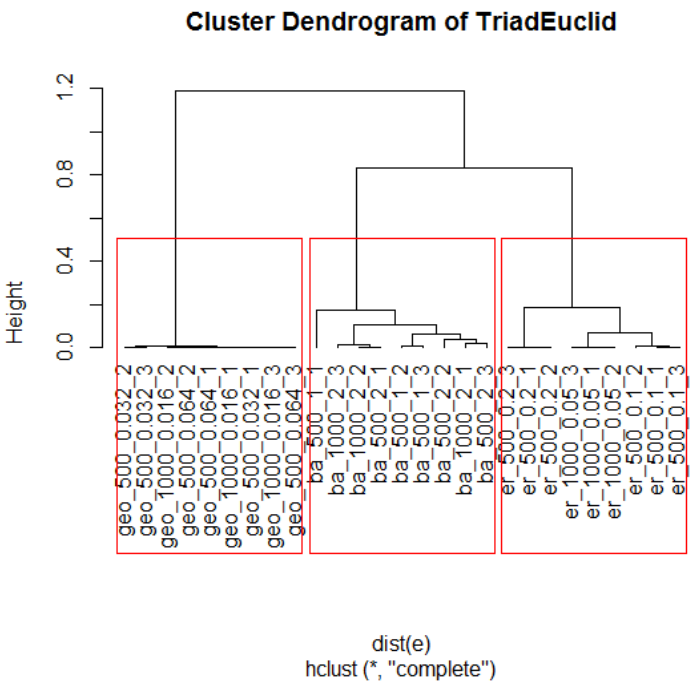

Figure 3: The tree shows the clustering result of TriadEuclid. There are three red boxes and each box represents one cluster.

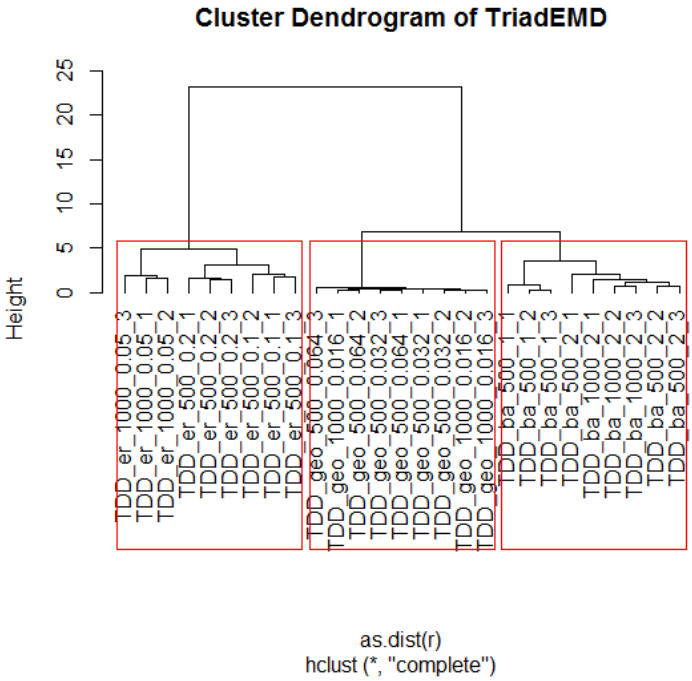

Figure 4: The tree shows the clustering result of TriadEMD. There are three red boxes and each box represents one cluster.

4 Signature triads for synthetic data

Table 2 gives the signature triads for the synthetic data set. Only the signature triads for the geometric random graphs show any variability.

Table 3 shows the signature orbits for the synthetic data set. There is considerably more variability than in the signature triads.

Table 2: Case Study I: Signature Triads

	Signature Triads
ba_1000_2_1	3
ba_1000_2_2	3
ba_1000_2_3	3
ba_500_1_1	3
ba_500_1_2	3
ba_500_1_3	3
ba_500_2_1	3
ba_500_2_2	3
ba_500_2_3	3
er_1000_0.05_1	10
er_1000_0.05_2	10
er_1000_0.05_3	10
er_500_0.1_1	-15
er_500_0.1_2	-15
er_500_0.1_3	-15
er_500_0.2_1	-15
er_500_0.2_2	-15
er_500_0.2_3	-15
geo_1000_0.016_1	9
geo_1000_0.016_2	15
geo_1000_0.016_3	15
geo_500_0.032_1	9
geo_500_0.032_2	15
geo_500_0.032_3	15
geo_500_0.064_1	0
geo_500_0.064_2	15
geo_500_0.064_3	15

Table 3: Case Study I: Signature Orbits

	Signature Orbit
ba_1000_2_1	12
ba_1000_2_2	12
ba_1000_2_3	12
ba_500_1_1	6
ba_500_1_2	19
ba_500_1_3	19
ba_500_2_1	12
ba_500_2_2	12
ba_500_2_3	19
er_1000_0.05_1	16
er_1000_0.05_2	1
er_1000_0.05_3	4
er_500_0.1_1	13
er_500_0.1_2	19
er_500_0.1_3	11
er_500_0.2_1	19
er_500_0.2_2	26
er_500_0.2_3	26
geo_1000_0.016_1	19
geo_1000_0.016_2	4
geo_1000_0.016_3	4
geo_500_0.032_1	12
geo_500_0.032_2	4
geo_500_0.032_3	4
geo_500_0.064_1	12
geo_500_0.064_2	19
geo_500_0.064_3	4

5 Details for Case Study II

The basic information about the networks in this case study is given in Table 4. Table 5 gives the clustering results. In Table 6, the clusters for the NNscore and N2Nscore calculations are given. The second column is the assumed cluster of each network, the third column is the cluster the nearest neighbour of each network, and the last column is the cluster of the 2nd nearest neighbour of each network.

The signature triads are shown in Table 7. For example, Triad 3 refers to a hierarchical structure with two reciprocal edges, whereas Triad 10 describes a directed flow, and Triad 15 is a reciprocal triad. The Triad No. is positive if it is over represented and negative if it is under represented. Table 8 gives the signature orbits.

Network	Nodes	Edges	Density	Assumed Cluster
Amazon0302.txt	262111	1234877	1.80e-05	1
Amazon0312.txt	400727	3200440	1.99e-05	1
Amazon0505.txt	410236	3356824	1.99e-05	1
Amazon0601.txt	403394	3387388	2.08e-05	1
Cit-HepPh.txt	34546	421578	3.53e-04	2
google_100129275726588145876.edges	1650	166292	6.11e-02	3
google_100329698645326486178.edges	2213	93510	1.91e-02	3
google_100466178325794757407.edges	344	4178	3.54e-02	3
google_100500197140377336562.edges	638	16043	3.95e-02	3
google_100518419853963396365.edges	326	10297	9.72e-02	3
google_100521671383026672718.edges	521	19847	7.32e-02	3
p2p-Gnutella04.txt	10876	39994	3.38e-04	4
p2p-Gnutella05.txt	8846	31839	4.07e-04	4
p2p-Gnutella06.txt	8717	31525	4.15e-04	4
p2p-Gnutella08.txt	6301	20777	5.23e-04	4
p2p-Gnutella09.txt	8114	26013	3.95e-04	4
p2p-Gnutella24.txt	26518	65369	9.30e-05	4
Slashdot0811.txt	77360	905468	1.51e-04	5
Slashdot0902.txt	82168	948464	1.40e-04	5
twitter_12831.edges	236	2478	4.47e-02	6
twitter_356963.edges	126	705	4.48e-02	6
twitter_428333.edges	65	1242	2.99e-01	6
twitter_612473.edges	76	1195	2.10e-01	6
twitter_613313.edges	88	968	1.26e-01	6
twitter_623623.edges	106	521	4.68e-02	6
twitter_629863.edges	159	1099	4.37e-02	6
twitter_734493.edges	9	54	7.50e-01	6
twitter_742143.edges	57	747	2.34e-01	6
twitter_745823.edges	242	6320	1.08e-01	6
twitter_759679.edges	115	782	5.96e-02	6
twitter_78813.edges	231	2861	5.38e-02	6

Table 4: The data set of different types of networks, and their assumed cluster

Network	Clustering Result
Amazon0302.txt	1
Amazon0312.txt	1
Amazon0505.txt	1
Amazon0601.txt	1
Cit-HepPh.txt	2
google_100129275726588145876.edges	3
google_100329698645326486178.edges	3
google_100466178325794757407.edges	3
google_100500197140377336562.edges	3
google_100518419853963396365.edges	4
google_100521671383026672718.edges	3
p2p-Gnutella04.txt	2
p2p-Gnutella05.txt	2
p2p-Gnutella06.txt	2
p2p-Gnutella08.txt	2
p2p-Gnutella09.txt	2
p2p-Gnutella24.txt	2
Slashdot0811.txt	5
Slashdot0902.txt	5
twitter_12831.edges	4
twitter_356963.edges	4
twitter_428333.edges	4
twitter_612473.edges	4
twitter_613313.edges	4
twitter_623623.edges	3
twitter_629863.edges	4
twitter_734493.edges	6
twitter_742143.edges	4
twitter_745823.edges	4
twitter_759679.edges	4
twitter_78813.edges	4

Table 5: Hierarchical Clustering Result for $TriadEuclid$

Table 6: NN scores and N2N scores for *TriadEuclid*

Network	Cluster	Cluster of NN	Cluster of 2nd NN
Amazon0302.txt	1	1	1
Amazon0312.txt	1	1	1
Amazon0505.txt	1	1	1
Amazon0601.txt	1	1	1
google_100129275726588145876.edges	3	3	3
google_100329698645326486178.edges	3	6	3
google_100466178325794757407.edges	3	3	6
google_100500197140377336562.edges	3	3	3
google_100518419853963396365.edges	3	6	6
google_100521671383026672718.edges	3	3	3
p2p-Gnutella04.txt	4	4	4
p2p-Gnutella05.txt	4	4	4
p2p-Gnutella06.txt	4	4	4
p2p-Gnutella08.txt	4	4	4
p2p-Gnutella09.txt	4	4	4
p2p-Gnutella24.txt	4	4	4
twitter_12831.edges	6	6	6
twitter_356963.edges	6	6	6
twitter_428333.edges	6	6	6
twitter_612473.edges	6	6	3
twitter_613313.edges	6	6	6
twitter_623623.edges	6	3	3
twitter_629863.edges	6	6	6
twitter_734493.edges	6	6	6
twitter_742143.edges	6	6	3
twitter_745823.edges	6	6	6
twitter_759679.edges	6	6	6
twitter_78813.edges	6	6	6

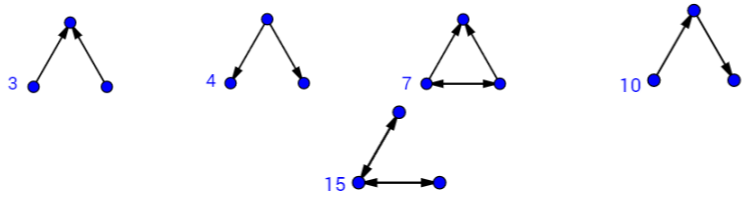

Figure 5: The signature triads for Case Study II

Table 7: Signature Triads for *TriadEuclid* of Different Networks. The Triad No. is positive if it is over-represented and negative if it is under-represented.

Network	Signature Triads No.
Amazon0302.txt	3
Amazon0312.txt	3
Amazon0505.txt	3
Amazon0601.txt	3
Cit-HepPh.txt	3
google100129275726588145876.edges	4
google_100329698645326486178.edges	3
google_100466178325794757407.edges	4
google_100500197140377336562.edges	4
google_100518419853963396365.edges	4
google_100521671383026672718.edges	4
p2p-Gnutella04.txt	10
p2p-Gnutella05.txt	3
p2p-Gnutella06.txt	10
p2p-Gnutella08.txt	3
p2p-Gnutella09.txt	3
p2p-Gnutella24.txt	4
Slashdot0811.txt	15
Slashdot0902.txt	15
twitter_12831.edges	3
twitter_356963.edges	15
twitter_428333.edges	-3
twitter_612473.edges	-3,4
twitter_613313.edges	-4
twitter_623623.edges	4
twitter_629863.edges	-15
twitter_734493.edges	7
twitter_742143.edges	-3
twitter_745823.edges	-3
twitter_759679.edges	-3
twitter_78813.edges	/

Table 8: Case Study II: Signature Orbits

	Signature Orbits
Amazon0302.txt	15
Amazon0312.txt	15
Amazon0505.txt	15
Amazon0601.txt	15
Cit-HepPh.txt	19
google_100129275726588145876.edges	13
google_100329698645326486178.edges	13
google_100466178325794757407.edges	13
google_100500197140377336562.edges	19
google_100518419853963396365.edges	14
google_100521671383026672718.edges	13
p2p-Gnutella04.txt	13
p2p-Gnutella05.txt	13
p2p-Gnutella06.txt	13
p2p-Gnutella08.txt	13
p2p-Gnutella09.txt	13
p2p-Gnutella24.txt	13
Slashdot0811.txt	12
Slashdot0902.txt	6
twitter_12831.edges	19
twitter_356963.edges	11
twitter_428333.edges	13
twitter_612473.edges	14
twitter_613313.edges	19
twitter_623623.edges	13
twitter_629863.edges	19
twitter_734493.edges	24
twitter_742143.edges	28
twitter_745823.edges	11
twitter_759679.edges	19
twitter_78813.edges	9

6 Details for Case Study III

The basic information of the World trade networks is shown in Table 9. The number of nodes is relatively small but the trade networks are generally of the same size. All of the networks are fairly dense, with lowest density 0.284.

Figures 6 and 7 show the heatmaps of the multi-layer world trade network comparisons, using $TriadEuclid$ and $TriadEMD$, respectively.

Table 9: The World Trade Network

Year	Nodes	Edges	Density
1962	154	8178	0.347
1963	156	8983	0.372
1964	158	9341	0.377
1965	157	9743	0.398
1966	157	10112	0.413
1967	157	10354	0.423
1968	156	10263	0.424
1969	158	10568	0.426
1970	162	11683	0.448
1971	162	11803	0.453
1972	162	12118	0.465
1973	163	12339	0.467
1974	162	12830	0.492
1975	163	12978	0.491
1976	163	12441	0.471
1977	163	11806	0.447
1978	162	9503	0.364
1979	163	13034	0.494
1980	163	12411	0.470
1981	163	12166	0.461
1982	162	10962	0.420
1983	162	9412	0.361
1984	168	7962	0.284
1985	170	8164	0.284
1986	171	8381	0.288
1987	170	8654	0.301
1988	168	8844	0.315
1989	173	9058	0.304
1990	174	9184	0.305
1991	172	9231	0.314
1992	188	10159	0.289
1993	188	10642	0.303
1994	187	10937	0.314
1995	186	11372	0.330
1996	187	11877	0.341
1997	190	12131	0.338
1998	188	12197	0.347
1999	189	12308	0.346
2000	190	11968	0.333

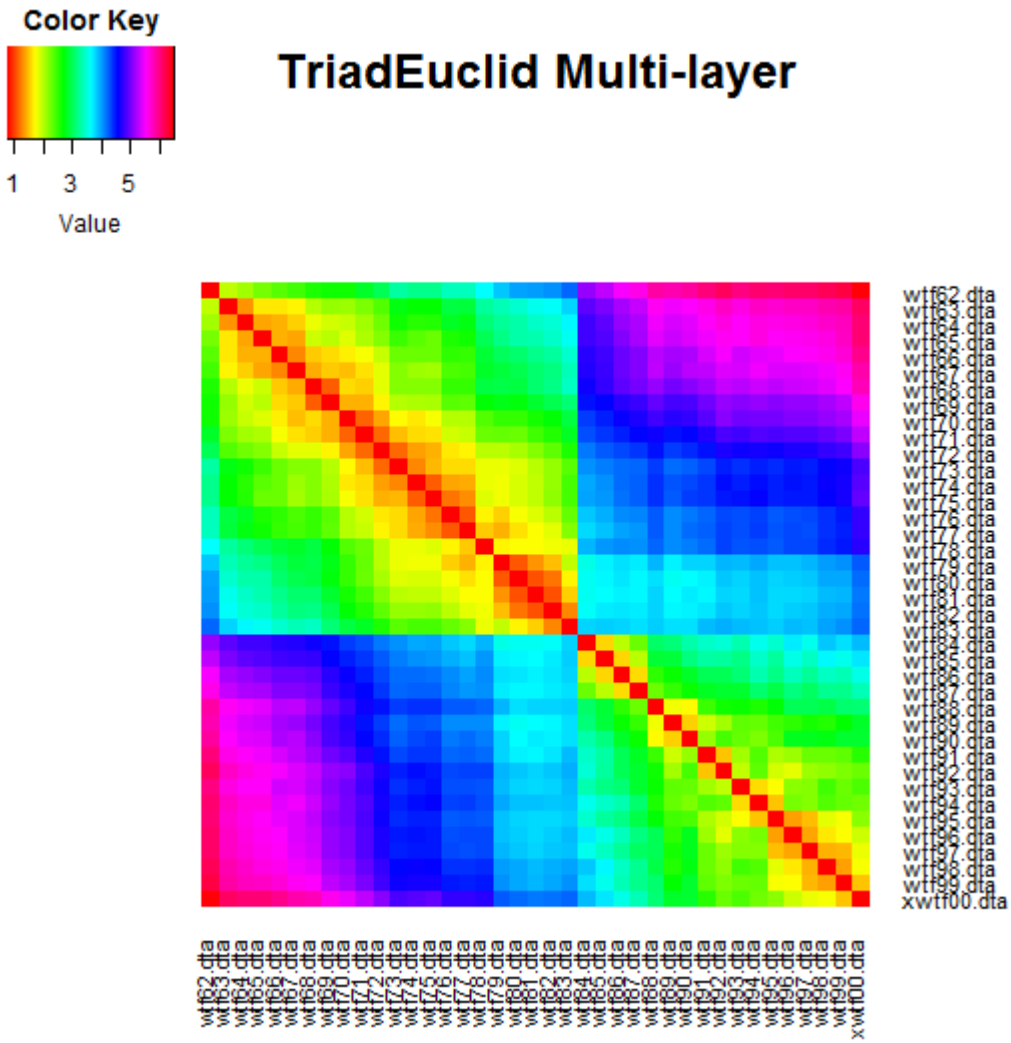

Figure 6: Heatmap of the Multi-layer World Trade Network using *TriadEuclid*

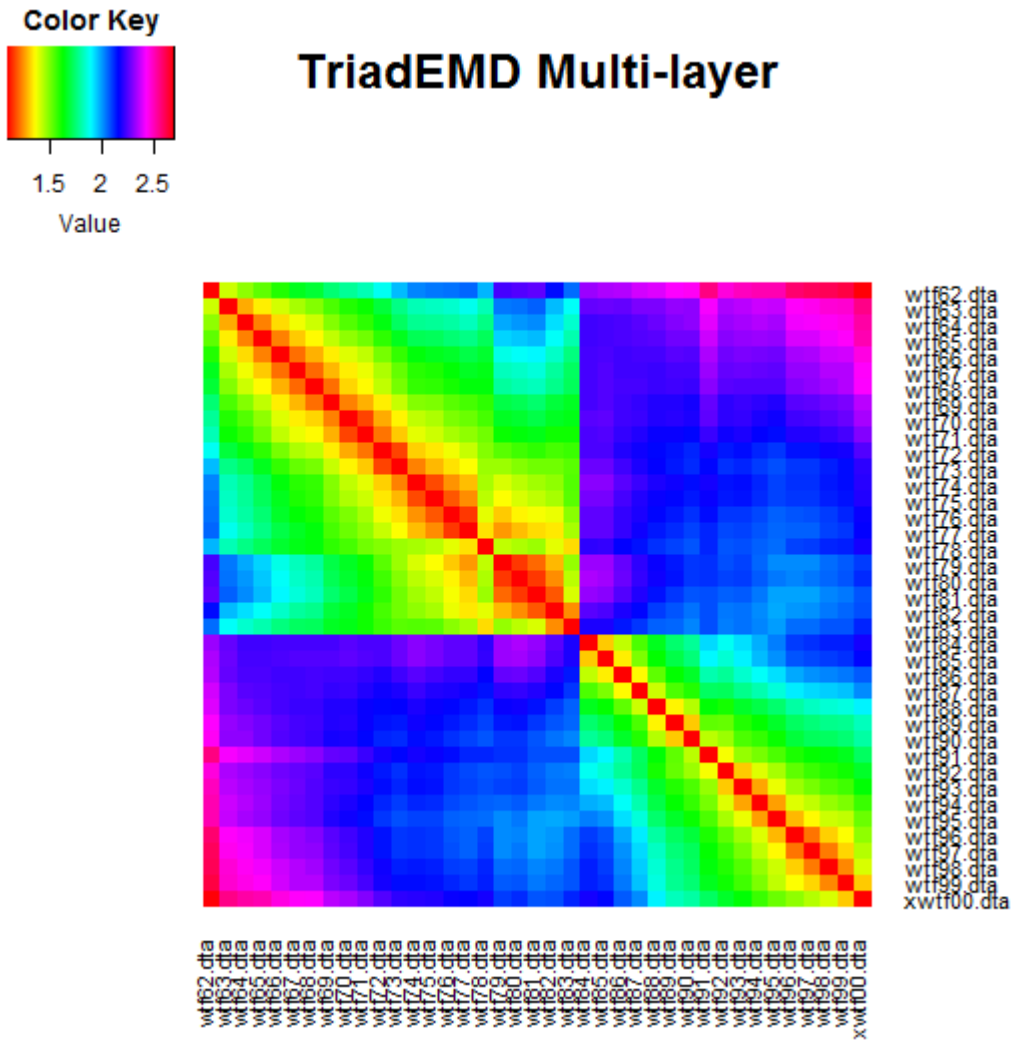

Figure 7: Heatmap of the Multi-layer World Trade Network using *TriadEMD*