American Journal of Applied Sciences 10 (7): 646-653, 2013
ISSN: 1546-9239

doi:10.3844/ajassp.2013.646.653 Published Online 10 (7) 2013 (http://www.thescipub.com/ajas.toc)

NETWORK TOPOLOGICAL PROPERTY OF ENGLISH DIALECTS SIMILARITY: A ROBUST FILTER APPROACH

Maman Abdurachman Djauhari and Gan Siew Lee

Department of Mathematical Sciences,
Faculty of Science, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Received 2013-01-28, Revised 2013-05-08; Accepted 2013-06-17

ABSTRACT

We show that the common approach to use of Minimum Spanning Tree (MST) to filter the information contained in a complex system of English dialects similarity is not robust. Later on we propose a robust filter based on the forest of all possible MSTs. To illustrate the advantages of this filter, Morgan's data on English dialects similarity is analyzed and promising results are reported.

Keywords: Centrality Measure, Minimum Spanning Tree, Networks Analysis, Single Linkage, Sub-Dominant Ultrametric

1. INTRODUCTION

The University of Leeds, as reported in Morgan (1981), has conducted a survey of English dialects that involved selecting over 300 English villages and interviewing carefully the chosen individuals from those villages to assess, amongst other things, their vocabulary. In his work, Morgan has focused his analysis on 25 East Midland villages, (**Fig. 1**) where a representative set of 60 items was chosen and for each pair of those villages, the percentage of the items for which the same word was used, was evaluated. Then, he considered these percentages as measures of dialect similarity. The results in the form of a similarity matrix is given in Morgan (1981) **Table 1**.

That similarity matrix can also be considered as a numerical summary of a complex system representing the 25 villages and their interrelationships in the form of network among villages. To analyze that network, Morgan used Minimum Spanning Tree (MST) and Sub-Dominant Ultrametric (SDU) as the tools. MST is used to filter the information contained in the network while SDU is to conduct Single Linkage Cluster Analysis (SLCA). Thus, if SLCA is used to illustrate the history of how those villages are clustered in the form of hierarchical tree, MST is to reveal features missed by the hierarchical tree. The details of the results and discussions can be found in Seber (2009). This shows the important roles of MST and SDU in extracting the information in the complex structure of the interrelationships among villages in terms of English dialects similarity. If SDU is to study the dialectical taxonomy of the villages, MST is to understand the topological properties of them. From the literature of networks analysis we learn that, nowadays, MST and SDU have become indispensible tools not only in linguistics but also in many areas of scientific investigation such as, for example, complex system, econophysics, financial time series, politics, portfolio optimization, social network and stocks market.

Due to the important roles of those tools in filtering the important information contained in any complex system, in this study we show that the use of MST might lead to non-robust information about the topological properties of the villages. Therefore, the interpretation of English dialects features will be misleading. To overcome this limitation, we propose a new filter which will give robust information and thus dialects features will be well described.

Corresponding Author: Maman Abdurachman Djauhari, Department of Mathematical Sciences, Faculty of Science, University Teknologi Malaysia, Johor Bahru, Malaysia Tel: +60-17-52-00795 Fax: +607-55-66162

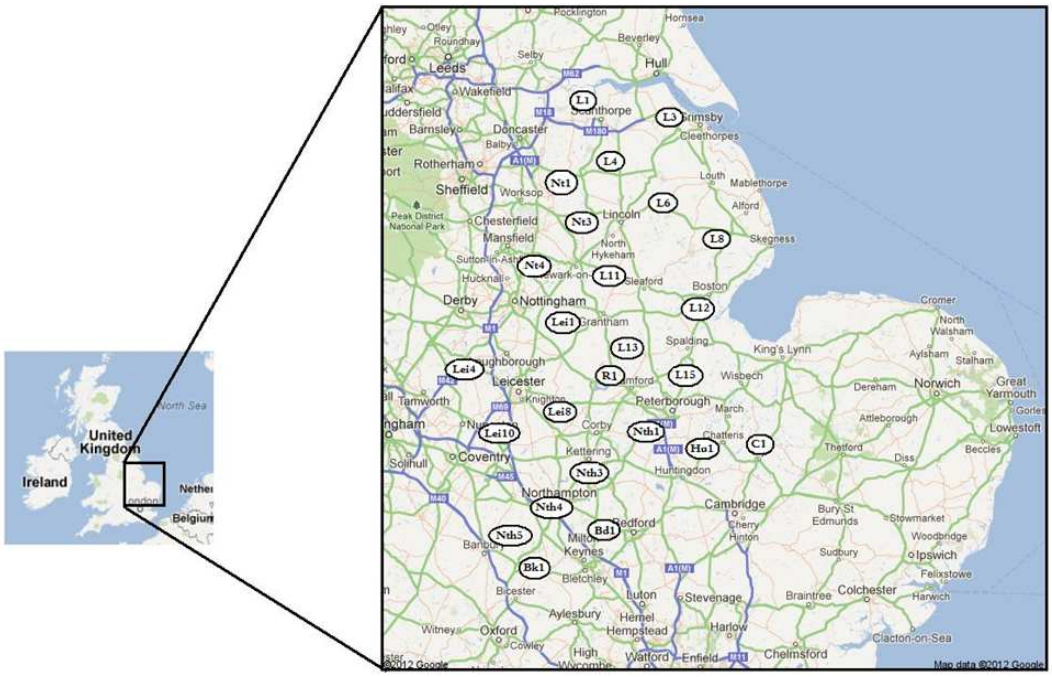

Fig. 1. Research area in east midland

$$S = \begin{pmatrix} 100 & 95 & 85 & 90 & 95 \\ 95 & 100 & 85 & 80 & 15 \\ 85 & 85 & 100 & 70 & 75 \\ 90 & 80 & 70 & 100 & 90 \\ 95 & 15 & 75 & 90 & 100 \end{pmatrix}$$

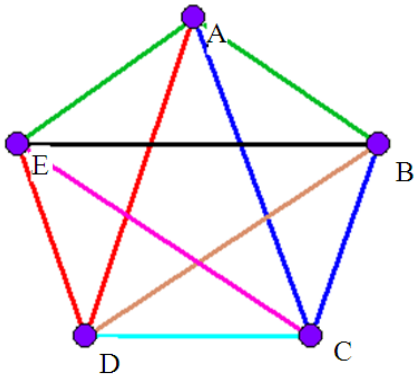

Fig. 2. Network among villages

We begin our discussion with the evidence that the use of MST to filter the information from English dialects similarity networks might give misleading information. Therefore, instead of using MST, we propose to use the forest of all possible MSTs as a robust filter.

This is delivered in the study followed by an algorithm to find the proposed filter. An analysis of English dialects similarity based on the proposed filter is reported and we find the advantages of the proposed method by comparing the results with those given by MST. Concluding remarks close the presentation.

2. NON-ROBUST OF MST-BASED FILTER

Consider the following hypothetical dialects similarity among five villages A, B, C, D and E represented in the form of similarity matrix S and network among similarities in **Fig. 2**.

Because of ties between the similarities in S, the MST in that network is not unique. This conclusion comes from the property that in any network, the MST is unique if and only if all elements of the corresponding similarity matrix are different to each other (Graham and Hell, 1985).

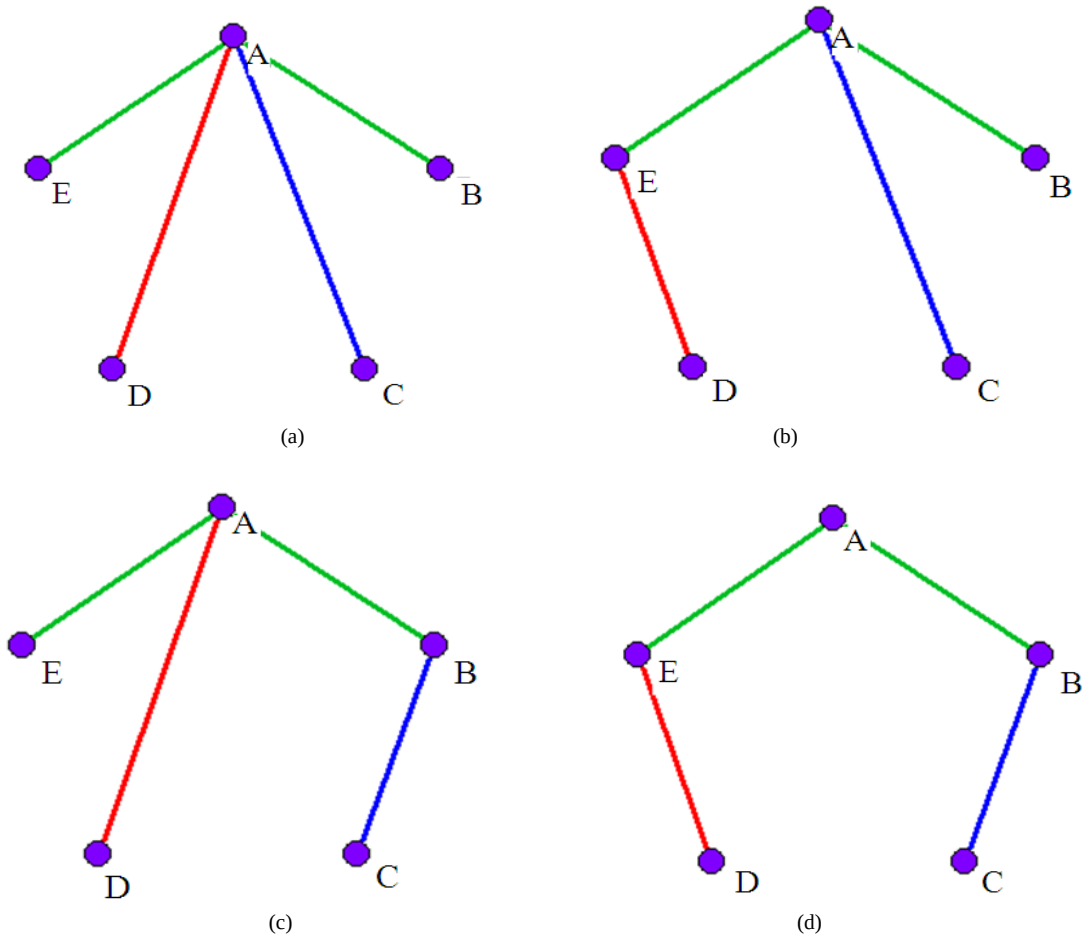

Fig. 3. All possible minimal spanning trees

Although Morgan (1981) mentioned about this property, however, he does not specify further how to handle this situation.

In that network there are four possible MSTs that could be given by any algorithm.

They are presented in **Fig 3a-d**. To the knowledge of the authors, the algorithms to construct MST, even the most popular and widely used algorithms such as Kruskal's algorithm or Prim's algorithm, only provide us with one single MST among all possible MSTs. The output of those algorithms depends on the data structure stored. Therefore, if **Fig. 2** represents dialect similarity, we can imagine how different the interpretation of dialects similarity based on the MST in **Fig 3a** compared to that given by **Fig 3d**. This shows the nonrobustness of the MST-based filter when the network contains more than one MST. To overcome the situation, in what follows we propose a robust filter.

3. PROPOSED ROBUST FILTER

Instead of using MST, in this study we propose to use the forest of all MSTs (or briefly the 'forest') to filter the information in a network among dialects similarity. If MST might not be unique in the network, the forest is unique. In other words, if one works based on MST, there might be many different possible network topologies that can be used as the filtered network. However, there is only one network topology if one works with the forest. That uniqueness guarantees the robustness of the filtered information provided by the forest. As an example, the forest of the network in **Fig. 2** is represented in **Fig. 4**.

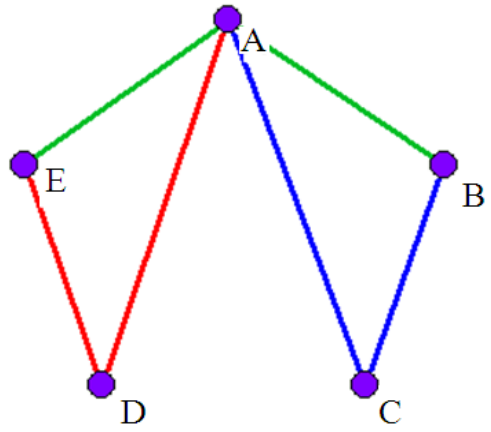

Fig. 4. The forest of all minimal spanning trees

In order to determine the forest in any network, an algorithm is provided. Let E be a set of n villages in a study of dialects similarity. We denote S the similarity matrix of size n rows and n columns representing the interrelationships among villages; the i-th row and j-th column element of S, denoted by s(i, j), is the similarity between the i-th and j-th villages. Thus, S is a symmetric matrix and by definition all diagonal elements are equal to 100%. If D is a matrix where d(i, j) = 100% - s(i, j) for all i, j = 1, 2, …, n, then D is the dissimilarity matrix associated to S. In practice, we can choose to work with S or with D. The results are the same. However, in what follows, our discussion will be based on D.

In current practice, villages and their interrelationships are considered as a complex system. Morgan (1981) and Seber (2009) for practical details. On the other hand, Tumminello *et al*. (2005) for general case, it is common to filter the information contained in complex system and to conduct analysis based on the filtered information. Therefore, if D contains one unique MST, the filtered information provided by MST is robust in the sense that there is no other source of filtered information that can be used. However, in practice, often D contains more than one MST. In this case, as we have mentioned in the previous paragraph, the use of MST might be misleading. This motivates us to use the forest as a robust filter. We will see later the advantages of the forest as a robust filter. For practical purpose, we develop an algorithm to construct that forest using fuzzy relation approach. This approach allows us to see the properties of D.

Let us consider k times minmax transitive operation '*' on D and itself, Djauhari (2012):

$$D^{*k} D * D^{*k-1} \text{ for all } k = 2, 3, \ldots$$

where, $D^{*1} = D$, the membership function D^{*k} in D^{*k} is defined by Equation (1):

$$d^{*k}(i,j) = \bigwedge_{m=1}^{n} \{d(i,m) \vee d^{*(k-1)}(m,j)\} \quad (1)$$

and $a \times b = \min\{a, b\}$ and $a \times b = \max\{a, b\}$ for all real numbers a and b. Since, D is symmetric and has anti-reflexive fuzzy relation with d as the membership function, the sequence D, D^{*2}, D^{*3},…, D^{*k}… is monotone decreasing, i.e.,:

$$\ldots \subseteq D^{*k} \subseteq \ldots \subseteq D^{*3} \subseteq D^{*2} \subseteq D$$

Djauhari (2012) for the proof. This property is very important because it can simplify the computation of the SDU of D and thus the single linkage in SLCA. We recall that SDU can be numerically represented as the minimax transitive closure D^+ of D, where:

$$D^+ = D \cap D^{*2} \cap D^{*3} \cap \ldots\ldots \cap D^{*k}$$

For any integer k. Djauhari (2012) for the details. Since the sequence D, D^{*2}, D^{*3},…, D^{*k}… is monotone decreasing, we simply have that SDU is $D^+ = D^{*k}$.

4. PROPOSED ALGORITHM

The SDU of D is unique but the MST of D might be not. If MST is unique, the forest consists of one single MST only. To construct the forest, let Δ be a fuzzy relation where its membership function δ is defined by Equation (2):

$$\delta(i,j) = \begin{cases} 1; d(i,j) - d^+(i,j) = 0 \text{ and } i \neq j \\ 0; d(i,j) - d^+(i,j) \neq 0 \text{ or } i = j \end{cases} \quad (2)$$

And d^+ is the membership function of D^+. In matrix form:

$$\Delta = \begin{pmatrix} \delta(1,1) & \delta(1,2) & \ldots & \delta(1,n) \\ \delta(1,2) & \delta(2,2) & \ldots & \delta(2,n) \\ M & M & O & M \\ \delta(n,1) & \delta(n,2) & L & \delta(n,n) \end{pmatrix}$$

Then, Djauhari (2012) for the proof, Δ is the adjacency matrix that corresponds to the forest. Thus, the forest is defined by all pairs (i, j) where i > j and δ(i, j) = 1. Furthermore, by observing the number N of those pairs, we obtain a necessary and sufficient condition for the uniqueness of MST in D.

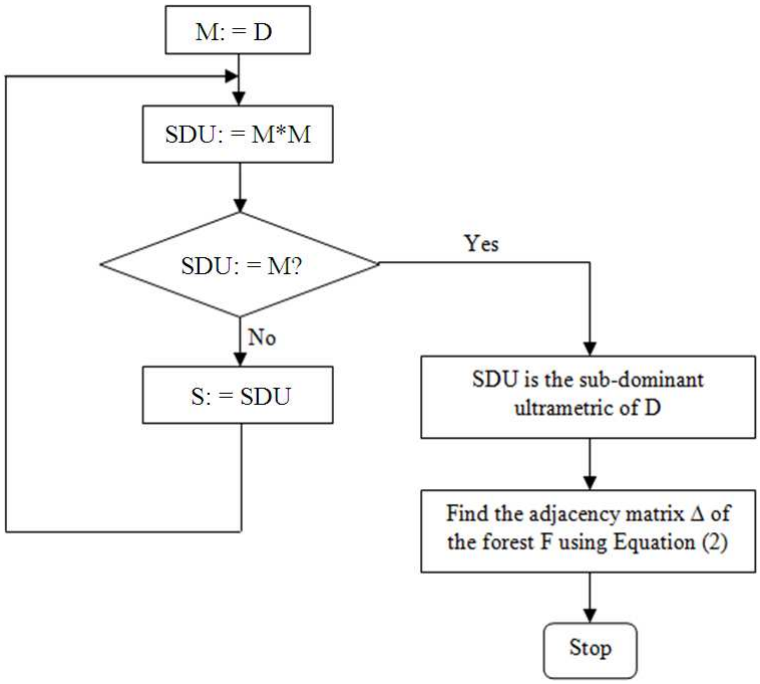

Fig. 5. Flow chart of SDU and forest construction

The MST in D is unique if and only if N = (n-1). This property can be used to check whether the MST in the network under study is unique or not.

Based on the property of SDU and Equation (2), the flow chart of the construction of SDU and the forest is presented in **Fig. 5**. This flow chart leads us to propose the following algorithm to obtain the forest. In this algorithm D and Δ are considered to be matrices of size (n×n).

Step1: Let k = 2,
Step2: Compute D^{*k} where $D*D^{*(k-1)}$ is a matrix multiplication in the usual sense but multiplication and summation of two real numbers a and b are defined as max{a, b} and min{a, b}, respectively,
Step3: If $D^{*k} = D^{*(k-1)}$, then the SDU of D is D^{*k} and go to Step 4. Otherwise, let k: = k + 1 and then go back to Step 2,
Step4: Compute Δ as defined in (2). Then Δ is the adjacency matrix representing the forest.

To speed up the convergence of that algorithm, instead of computing D^{*2}, D^{*3}, D^{*4}, …, we compute directly D^{*2}, D^{*4}, D^{*8}, … . In this case, the computation process is stopped at the *K*-th iteration if $D^{*2^K} = D^{*2^{(K-1)}}$ and the number of iterations needed is $K < \frac{\ln(n)}{\ln(2)}$.

5. ANALYSIS OF ENGLISH DIALECT SIMILARITIES

Consider again the network among villages representing the complex system of 25 villages where their interrelationships is numerically summarized as the similarity matrix S given in Morgan (1981) **Table 1**. To analyze that network we use the dissimilarity matrix D where d(i, j) = 100 %-s(i, j) for all i, j = 1, 2, …, n = 25.

We start by studying the topological properties of villages based on MST in terms of their centrality measures. These measures will help us to have a better understanding about the social power of each village relative to the others in the formation of the network. Later on those properties will be studied based on the forest and see the difference.

5.1. Information Filtering

Figure 6 below represents the network topology of the 25 villages based on an MST in accordance with their geographical location Morgan (1981) and Seber (2009).

This figure does not only reveal a northsouth dichotomy of the villages but also a westeast dichotomy.

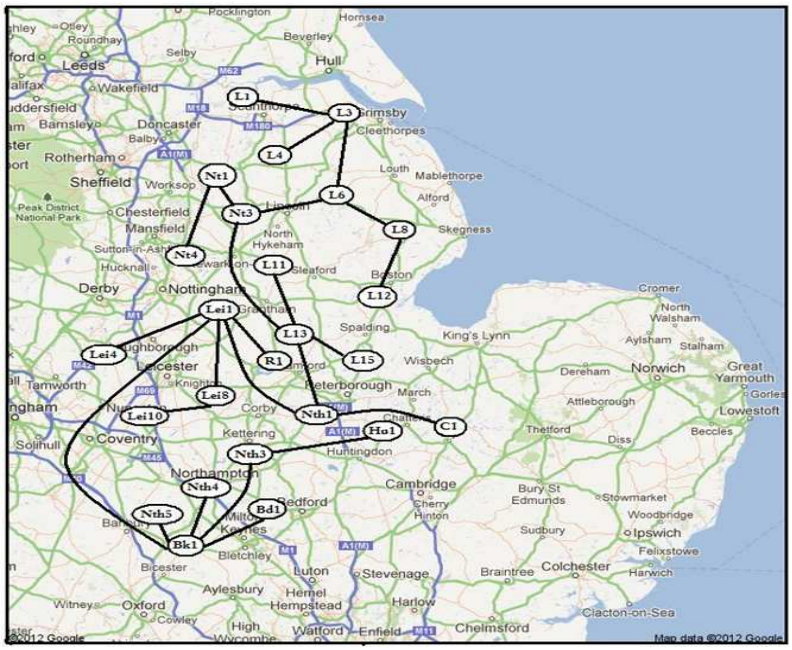

Fig. 6. Network topology based on an MST

It also determines the social power of each village with respect to the others in terms of its centrality measures such as degree, betweenness, closeness and eigenvector centralities. Degree centrality refers to the number of links that each village has. The more links a village has, the more power it may have. However, having the same degree does not necessarily make villages equally important. On the other hand, betweenness refers to the positional advantage of villages which fall on the shortest geodesic pathway between other pairs of villages (Freeman, 1979). Villages with high betweenness scores are those that act as coordinating the influence of a village's dialect behavior to the others. The third measure of centrality, i.e., closeness centrality, is to identify the villages that are able to influence other villages at shorter path lengths, or that are more influenced by other villages at shorter path lengths (Bonacich, 1987). Finally, eigenvector centrality measure is to know the villages that have links to other powerful villages (Borgatti, 2005).

In **Fig. 6**, there are 14 leaves. They are the worst villages that receive influence from others. They receive the smallest scores in all measures. The remaining 11 villages have various scores in those measures. According to the four measures mentioned above, the list of these villages ordered from the strongest until the weakest is given in **Table 1**.

Table 1. The list of villages ordered from the strongest to the weakest score in the centrality measures issued from MST

No.	Degree	Betweenness	Closeness	Eigenvector
1	Lei1	Lei1	Nth1	Lei1
2	Bk1	L13	L13	Bk1
3	L13	Nth1	Lei1	Nth1
4	Nt3	Nt3	Nt3	L13
5	L3	Bk1	Bk1	Lei8
6	L6	L6	L6	Nth3
7	Nth1	L3	Lei8	Nt3
8	Nt1	Nt1	Nt1	L6
9	L8	L8	Nth3	Nt1
10	Lei8	Lei8	L3	L3
11	Nth3	Nth3	L8	L8

It is interesting to note that Lei1 (Leicestershire) has the highest score in all measures except "Closeness" and that Nth1 (Nottinghamshire 1) is the closest village to all other villages. In the next paragraph we show that the network of dialects similarity among villages contains more than one MST, as can be seen in **Fig. 7**. Therefore, the information contained in the network topology in **Fig. 6** is not robust.

5.2. Robust Topological Properties

Morgan (1981), MST is used to reveal features missed by SLCA. Although he mentions the nonuniqueness of MST, he does not explore its consequence. In **Fig. 7** we present the network topology of villages based on the forest.

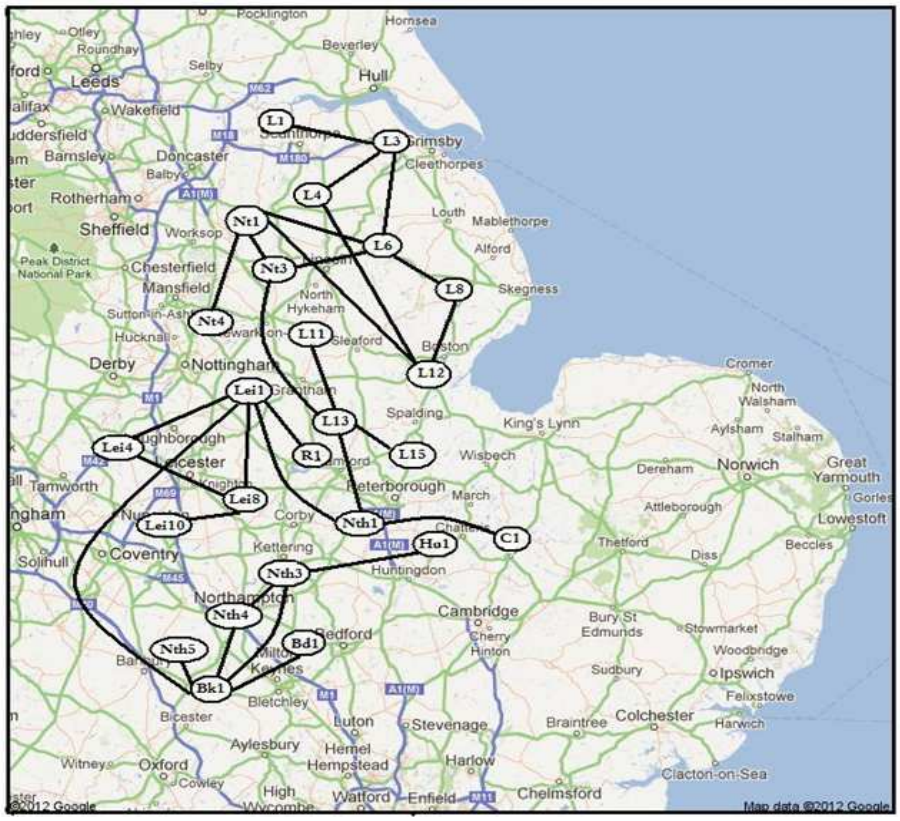

Fig. 7. Network topology based on the forest

Table 2. The list of villages ordered from the strongest to the weakest score in the centrality measures issued from the forest

No.	Degree	Betweenness	Closeness	Eigenvector
1	Lei1	Lei1	Nth1	Lei1
2	Bk1	L13	L13	L6
3	Nt1	Nth1	Lei1	Nt1
4	L6	Nt3	Nt3	Nt3
5	L13	Bk1	Bk1	Bk1
6	Nt3	L6	L6	L13
7	L3	Nt1	Nt1	L12
8	L12	L3	Lei8	Nth1
9	Lei8	Lei8	Lei4	Lei8
10	Nth1	Nth3	Nth3	L8
11	Nth3	L12	Nth4	Lei4
12	L4	L4	L3	L3
13	L8	L8	L12	Nth3
14	Lei4	Lei4	L8	Nth4
15	Nth4	Nth4	L4	L4

Unlike in **Fig. 6**, in **Fig. 7** we have 10 leaves. The remaining 15 villages have various scores in the four centrality measures. The list of these villages ordered from the strongest to the weakest score in influencing other villages is given in **Table 2**.

From this table we learn that the topological properties of the network among villages' similarities provided by the forest is totally different from those by MST except the position of Lei1 (Leicestershire) and Nth1 (Nottinghamshire 1). Interestingly:

- The forest consists of 24 MSTs. Thus, there are 24 possible network topologies based on MST and two different networks have different topological properties
- The similarity score of the pair (Nt1, L6) and that of (Nt1, Nt3) are the same, i.e., 71% as given in Morgan (1981), **Table 1**
- The pairs (Lei4, Lei1) and (Lei4, Lei8) also have the same similarity score of 63% as given in Morgan (1981), **Table 1**
- The similarity scores of (Nth8, Bk1) and (Nth8, Nth4) are also the same, i.e., 63% as given in Morgan (1981), **Table 1**

- L12 at the east links to L4 and Nt1 at the North West with the same amount of similarity as of L12 with L8 which is also at the east. Geographically, L4 and Nt1 are far from L12 compared to L8. Therefore, L12 has a special feature in the study of English dialects similarity
- Similar feature is also possessed by Nt1at the west which links also with L6 at the east

6. CONCLUSION

We show that MST as an information filter is not robust except when it is unique. To overcome that obstacle we propose to use the forest of all possible MSTs as a robust filter. For practical purpose, an algorithm to obtain the forest is provided. The study of English dialects similarity, initiated by Morgan (1981), using the forest is more advantageous than MST.

Among many advantages of the forest as information filter, in terms of degree centrality, can be mentioned here. By using MST, Morgan highlights Buckinghamshire 1, Leicestershire and Lincolnshire 13 as villages with four or more links. Actually, according to the forest, Nottinghamshire 1 and Lincolnshire 6 must receive similar consideration like those three villages. They also have four links. In terms of other centrality measures, the power of each village in influencing the others issued by the forest is totally different from that issued by MST except Lei1 and Nth1.

7. ACKNOWLEDGEMENT

The researcher are very grateful to the anonymous referees for their comments and suggestions that led to the final presentation of this study. They also gratefully acknowledge Government of Malaysia for the sponsorships, through Fundamental Research Grant Schemes vote number 4F013 and Research Universiti Teknologi Malaysia Grant vote number 02H18 and Universiti Teknologi Malaysia for the opportunity to do this research.

8. REFERENCES

Bonacich, P., 1987. Power and centrality: A family of measures. Am. J. Sociol., 92: 1170-1182. DOI: 10.2307/2780000

Borgatti, S.P., 2005. Centrality and network flow. Soc. Netw., 27: 55-71. DOI: 10.1016/j.socnet.2004.11.008

Djauhari, M.A., 2012. A robust filter in stock networks analysis. Physica A, 391: 5049-5057. DOI: 10.1016/j.physa.2012.05.060

Freeman, L.C., 1979. Centrality in networks: I. Conceptual clarification. Social Net., 1: 215-239. DOI: 10.1016/0378-8733(78)90021-7

Graham, R.L. and P. Hell, 1985. On the history of the minimum spanning tree problem. Annals History Comput., 7: 43-57. DOI: 10.1109/MAHC.1985.10011

Morgan, B.J.T., 1981. Three applications of methods of cluster-analysis. J. Royal Stat. Soc., 30: 205-223.

Seber, G.A.F., 2009. Multivariate Observations. 1st Edn., John Wiley and Sons, New York, ISBN-10: 0470317310, pp: 712.

Tumminello, M., T. Aste, T.D. Matteo and R.N. Mantegna, 2005. A tool for filtering information in complex systems. Proc. Natl. Acad. Sci. USA., 102: 10421-10426. DOI: 10.1073/pnas.0500298102