A Clustering Coefficient to Identify Important Nodes in Bipartite Networks

J. Liebig and A. Rao

School of Mathematical and Geospatial Sciences, RMIT University, Melbourne

Dated: June 24, 2014

Abstract

Bipartite networks have gained an increasing amount of attention over the past few years. Network measures in particular, have been the focus of this research as many of them cannot be directly applied to bipartite networks. The clustering coefficient is one measure that has been redefined recently to suit the analysis of bipartite networks. Building up on this definition, we propose a clustering coefficient that distinguishes between differently structured bipartite clusters. We use this measure to identify influential nodes in a given bipartite network. By comparing the global and local clustering coefficients, we assign a score to each node that indicates the extent to which it drives the clustering behaviour of the whole network. We demonstrate that our clustering coefficient is not only able to identify influential nodes, but gives new insights into a network’s structure.

1 Introduction

Networks model the interaction between the members of a system and arise in many different areas such as chemistry, epidemiology and sociology. A network consists of nodes that may be connected via edges. Facebook is an example of a social network, in which users are represented by nodes and an edge exists between two nodes if the corresponding people are Facebook friends. Insight and information about a network of interest can be gained using

1
a variety of network measures. Some examples of network measures include centrality measures, the clustering coefficient, etc. For a review on network measures see reference [10]. A Facebook friendship network may be thought of as a one-mode network of people who create connections to each other.

Here we are interested in a particular type of network that is very common in real life. For instance, Newman [21] investigates the structure of scientific collaboration networks. In [21] two authors are connected via an edge if they have written a scientific paper together. Similarly, Fortuna et al. [15] investigate a network of bats and trees, where a bat is linked to a tree if the bat roosted on that particular tree. In order to identify the most important trees, [15] considers the network that only consists of trees, by linking two trees if they were visited by the same bat, and applies centrality measures to the network. Both, [15] and [21], treat the network in question as a one-mode network. However, both networks are originally bipartite, like many other systems [17].

A bipartite network is a special type of network. The nodes of a bipartite network can be partitioned into two disjoint sets where nodes belonging to the same set cannot be connected through an edge. Bipartite networks are also called two-mode or affiliation networks. In [21], there are two types of nodes, authors and papers, where edges do not exist between nodes of the same type. In [15] the network of interest consists of bats and trees and an edge can only connect a bat to a tree. In both cases, in order to analyse the network, an approach, called one-mode projection or ‘conversion’ was used. The original bipartite network is projected down to a one-mode network by dropping one of the two node sets. Two nodes of the remaining node set are then connected if they share a neighbour in the bipartite network. Often, one-mode projection is necessary in order to analyse a bipartite network, mainly because most network measures cannot be directly applied to bipartite networks. However, many researchers agree that one-mode projection leads to loss of information. This issue has received increasing attention over the past few years and many measures have been redefined to suit the analysis of bipartite networks.

This paper focuses on one particular measure, the clustering coefficient, and it shows how this measure can be used to identify important and influential nodes within a network. In a one-mode network, where a connection through an edge is possible between any two nodes, the
clustering coefficient measures the concentration of 3-cycles [23]. One can measure the global clustering coefficient of a whole network or calculate the local clustering coefficient of a given vertex within a network. The clustering coefficient is an important measure as it shows how well the neighbours of a node are connected [3]. The definition of a bipartite network implies that no cycles of odd length can exist in such a network. It follows that the clustering coefficient cannot be directly applied to bipartite networks [24].

In the literature, several definitions for the clustering coefficient in a two-mode network can be found. A few examples are [20], [24], [26] and [32]. The majority of authors define the two-mode clustering coefficient to be the concentration of cycles of length 4. Opsahl’s definition [24] on the other hand differs from all previous definitions and defines the clustering coefficient in terms of cycles of length 6.

The clustering coefficient that is proposed in [24] and the original clustering coefficient for one-mode networks measure closure between three nodes and hence are concerned with small clusters. On the other hand, much work has been done in identifying larger clusters [2, 31, 22], sometimes referred to as communities or groups. The members of a group can further be divided into core and periphery nodes [12, 5], where core nodes play an important role within the network. Here, we do not aim to identify the different communities of a network, but to recognise influential nodes that may belong to different communities within the the network. The identification of influential nodes is crucial in many different areas. For instance, knowledge of influential nodes aids in terminating the spread of diseases. It may assist the spread of knowledge and information [18]. Important nodes can be identified by applying centrality measures or finding hubs. However, if the network has a community structure, centrality measures may only identify nodes within one community. Similarly, hubs may lie within the same community [33].

This paper proposes a clustering coefficient for bipartite networks that, besides distinguishing between differently structured clusters, also identifies important nodes that drive the clustering behaviour of the whole network. We apply our method to two real world data sets; the first being the popular southern women network [12] that has been analysed many times [16]. We show that the nodes identified by our method as influential, are indeed nodes that were classified as core nodes by previous analyses.

The rest of the paper is organised as follows: Section 3 identifies different types of 6-cycles and
shows that by distinguishing between these, differently structured clusters can be revealed. As the origin of clusters is important for the analysis, it is necessary to separate bipartite networks into different types, according to the way in which they develop over time. The proposed method of calculating the clustering coefficient is detailed in this section. Section 4 presents and discusses results that are obtained from two real world data sets. The results give insight into how clusters are structured and reveal the nodes that drive the formation of clusters and their particular structure.

2 The Bipartite Clustering Coefficient

Most existing methods for calculating the clustering coefficient in a bipartite network measure the concentration of squares (cycles of length 4) in the network of interest (see references [20], [26] and [32]), because a 4-cycle is the smallest possible cycle in a bipartite network and the clustering coefficient in a one-mode network measures the concentration of 3-cycles, the smallest possible cycles in this type of network.

A 4-cycle in a bipartite network shows that two nodes of the same type are connected twice via two nodes of the other type. For this reason and because the original clustering coefficient for one-mode networks measures closure between three nodes, Opsahl [24] chooses to define the clustering coefficient for bipartite networks in terms of paths of length 4 and cycles of length 6:

$$C^* = \frac{\text{closed 4-paths}}{4\text{-paths}} = \frac{\tau^*_\Delta}{\tau^*},$$

where $\tau^*_\Delta$ is the number of 4-paths and $\tau^*_i$ is the number of these 4-paths that are closed. A closed 4-path is equivalent to a cycle of length 6.

The local clustering coefficient for a node $v_i$ is given in [24] as follows:

$$C^*(i) = \frac{\tau^*_{i,\Delta}}{\tau^*_i},$$

where $\tau^*_{i,\Delta}$ is the number of 4-paths that are centred at node $v_i$ and $\tau^*_i$ is the number of these
Figure 1: There are two possible ways to connect three different active nodes to each other in a two-mode graph [24]. In the subgraph to the left, every active node is connected to the other active nodes via a different passive node, whereas in the subgraph to the right, all three active nodes are connected to each other via the same passive node.

4-paths that are closed.

Indeed, the bipartite clustering coefficient should measure closure between three nodes that belong to the same node set as it does in one-mode networks. Hence, the idea of triadic closure, presented in [24], is the obvious direction to follow. To better distinguish between the two node sets of a bipartite network, they are henceforth referred to as active and passive nodes. Active and passive nodes are represented by circles and squares respectively. Note that often the terms primary and secondary nodes are used in the literature to distinguish between the two node sets, but we use the terms active and passive to differentiate between the nodes that actively make the connections (the active nodes) and the nodes that are connected to (the passive nodes).

There are two subgraphs that form triangles when projected down to a one-mode network [24] (see Fig. 1). Figure 1a shows a 6-cycle which is considered to be a closed 4-path; a 4-path can be closed by connecting a new passive node to the two end nodes of a 4-path. Hence, a 4-path may be thought of as a possible 6-cycle. Figure 1b shows a 3-star which is also closed as the three active nodes are all connected to each other via the same passive node. As star subgraphs of any size are the reason for the count of triangles in a projected one-mode network to be higher than expected [24], the bipartite clustering coefficient, as defined by Eq. (1), does not consider a 3-star as closure between three active nodes. Only cycles of length 6 are considered as a closed connection between three nodes of the same type. We follow this convention in this paper.

There is an important difference between the one-mode and two-mode subgraphs involved in calculating the respective clustering coefficients. In a bipartite network, the nodes of a 6-cycle as
Figure 2: (a) A bipartite cycle of length 6 can have at most three additional edges, whereas a bipartite 4-path can have at most two. Additional edges are represented by dashed lines. Different numbers of additional links within the two subgraphs create different structures that have distinct meanings depending on the type of interaction that is modeled by the network. (b) In a one-mode network, no additional edges can be added to a triangle or a 2-path without creating multiple edges. Well as the nodes of a 4-path may be connected to each other by additional edges (see Fig. 2a), whereas in a one-mode network, additional edges between the nodes of a triangle or between the nodes of a 2-path cannot exist without producing multiple edges. Additional edges in the two-mode subgraphs give rise to different subgraphs with distinct meanings depending on the network. In this paper, these structures are considered separately.

3 Clusters with different structure

This section gives a different method for calculating the clustering coefficient that distinguishes between differently structured clusters, with the aim of identifying influential nodes in a given network. We start with some definitions.

Definition 3.1. Let $G$ be an undirected, unweighted bipartite graph without any multiple edges. The two disjoint node sets are denoted $V_1 = \{v_0, v_1, \ldots, v_m\}$ and $V_2 = \{w_0, w_1, \ldots, w_n\}$ respectively. $B(G)$ is the biadjacency matrix of $G$, such that

$$B(G) = [b_{ij}],$$

$$b_{ij} = \begin{cases} 
1 & \text{if node } v_i \text{ is connected to node } w_j \\
0 & \text{otherwise} 
\end{cases}$$

Definition 3.2. A cycle is a path that starts and ends at the same vertex. All other elements must be unique. A cycle of length $k$ is called a $k$-cycle and for a bipartite graph, may be written as $C_k = \{v_0, w_1, \ldots, w_{k-1}, v_0\}$. 

6
Note that a cycle can start and end at any node within the cycle. We consider this to be the same cycle as $C^k$. As this paper is concerned with undirected networks, the direction of $C^k$ does not matter and hence $C^k = \{v_0, w_1, \ldots, w_{k-1}, v_0\} = \{v_0, w_{k-1}, \ldots, w_1, v_0\}$.

It is possible to form different 6-cycles between the same six nodes by traversing different edges. The number of different 6-cycles depends on the number of edges that connect these six nodes to each other.

**Lemma 3.3.** Let $G$ be an undirected, unweighted bipartite graph and $B(G)$ its biadjacency matrix. Any 6-cycle $C^6$ in $G$ can be represented by a $3 \times 3$ submatrix $S$ of $B(G)$ where every row and every column of $S$ contains at least two 1s. $S$ can represent exactly one 6-cycle, exactly two 6-cycles or exactly six 6-cycles.

**Proof** Let $C^6 = \{v_i, w_r, v_j, w_s, v_k, w_t, v_i\}$ be a 6-cycle in $G$. Every node that is part of a 6-cycle must have degree at least 2 and exactly two of the incident edges must be part of the cycle [28]. All other edges are dropped from consideration. It follows that, any 6-cycle in $G$ can be represented by a $3 \times 3$ submatrix of $B(G)$ that contains at least two 1s in every row and every column. $C^6$ may then be represented by $S = \begin{bmatrix} 1 & b_{is} & 1 \\ 1 & 1 & b_{jt} \\ b_{kr} & 1 & 1 \end{bmatrix}$. The entries $b_{is}, b_{jt}$ and $b_{kr}$ are not part of $C^6$ and may equal 0 or 1. Note that since it does not matter which row represents which active node and which column represents which passive node, the rows and columns of $S$ can be rearranged to give $S = \begin{bmatrix} b_{is} & 1 & 1 \\ 1 & b_{jt} & 1 \\ 1 & 1 & b_{kr} \end{bmatrix}$, placing the entries that are not part of $C^6$ in the diagonal of the matrix.

Let $S = \begin{bmatrix} b_{is} & 1 & 1 \\ 1 & b_{jt} & 1 \\ 1 & 1 & b_{kr} \end{bmatrix}$ be a submatrix of $B(G)$. There are then four possibilities: 1. $S$ contains exactly three zero entries, 2. $S$ contains exactly two zero entries, 3. $S$ contains exactly one zero entry or 4. $S$ does not contain any zero entries. In order to find all 6-cycles that $S$ represents, all entries that are equal to 0 are dropped from consideration as they represent non-existing edges. Every node that is part of a 6-cycle has two incident edges that are also part of the cycle. Thus, if an entry of $S$ is dropped from consideration, all other entries in the
Figure 3: A $3 \times 3$ submatrix $S$ of $B(G)$ that contains at least two $1$s in every row and every column represents exactly one of the four cycles shown.

corresponding row and column must be part of the 6-cycle.

1. All entries that are equal to 0 are dropped. Then the entries in the corresponding row and column must be part of the 6-cycle and hence $S$ represents exactly one 6-cycle.

2. Without loss of generality, assume that $b_{is} = b_{jt} = 0$ and $b_{kr} = 1$. The two entries that are equal to 0 are dropped. Then the entries in the corresponding rows and columns must be part of the 6-cycle and hence $S$ represents exactly one 6-cycle.

3. Without loss of generality, assume that $b_{is} = 0$ and $b_{jt} = b_{kr} = 1$. The entry that is equal to 0 is dropped. Then the entries in the corresponding row and column must be part of the 6-cycle. Hence we can drop any one of the two remaining entries in the second column, thus giving
\[
\begin{pmatrix} 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2 \text{ 6-cycles.}
\]

4. Any one of the three entries in the first column is dropped from consideration. Then the entries in the corresponding row and column must be part of the 6-cycle. Then any one of the two remaining entries in the second column is dropped. Thus giving
\[
\begin{pmatrix} 3 \\ 1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = 6 \text{ 6-cycles.}
\]

Therefore $S$ can represent exactly one 6-cycle, exactly two 6-cycles or exactly six 6-cycles. □

The number of 6-cycles that are represented by $S$ depends on the number of entries that are equal to 0. If $S$ represents more than one 6-cycle, the different cycles involve the same nodes and therefore, the number of different 6-cycles represented by $S$ is irrelevant. $S$ corresponds to
Figure 4: Person a and person b attended event 1 that took place at time $t_0$. Person a and person c attended event 2 that took place at time $t_1$. Finally, person b and person c attended event 3 that took place at time $t_2$. At time $t_2$ the 6-cycle is complete and no connections within the cycle can be formed at a later point in time.

one of the subgraphs shown in Fig. 3. Our method distinguishes between these subgraphs and treats each as a 6-cycle of different type. Distinguishing between the different structures gives insight into how interconnected any set of three nodes of the same type can be. We call the structures shown in Fig. 3a, 3b, 3c and 3d an unconnected 6-cycle, a sparsely connected 6-cycle, a highly connected 6-cycle and a completely connected 6-cycle respectively.

Corollary 3.4. Let $G$ be an undirected, unweighted bipartite graph. Let $\gamma$ be the number of 6-cycles in $G$, $\gamma(0)$ the number of unconnected 6-cycles in $G$, $\gamma(1)$ the number of sparsely connected 6-cycles in $G$, $\gamma(2)$ the number of highly connected 6-cycles in $G$ and $\gamma(3)$ the number of completely connected 6-cycles in $G$. Then $\gamma = \gamma(0) + \gamma(1) + 2\gamma(2) + 6\gamma(3)$.

We develop four bipartite clustering coefficients, one for each type of 6-cycle, in order to gain new information about the network that is not revealed by previously defined clustering coefficients.

In order to calculate the clustering coefficient, we need to determine all possibilities by which the different types of 6-cycles can be formed. Most networks develop over a period of time and the way in which 6-cycles are formed depends on this process. Since data sets are often missing time stamps, our aim is to measure the clustering coefficient of a network at some point in time. Following we show why it is necessary to distinguish between different types of bipartite networks first. In some bipartite networks an active node can only connect to a particular passive node at a particular point in time. We call these networks time dependent networks. In a network that models the attendance of people at events, where each event takes place at a specific point in time.
time, it is impossible to form links within an existing 6-cycle. For instance, assume there exists a 6-cycle that is part of a network of people and events, see Fig. 1. Events 1, 2 and 3 take place at times $t_0, t_1$ and $t_2$ respectively. Figure 4 clearly shows that connections within 6-cycles cannot be formed at a later time $t_i$, as all three events have passed.

In networks where nodes are allowed to connect to passive nodes at any point in time, it is possible to build connections within an existing 6-cycle. A network of online forums and users, where forums are represented by passive nodes, is an example of such a network. The forums are accessible over a period of time and hence it is possible for connections to be formed within an existing 6-cycle. It is obvious that the origins of the different 6-cycles are distinct, depending on how the network develops. In this paper we only consider time dependent networks, as defined in the paragraph above.

As an example of a time dependent network we choose a popular, often analysed, data set that was collected by Davis, Gardner and Gardner and published in 1941 [12]. We refer to this network as the southern women network. This network consists of 18 women and 14 events. An edge between a woman and an event exists only if the woman attended the event. This bipartite network is clearly a time dependent network as the events take place at a certain time only and cannot be attended afterwards. Every event is associated with a different time $t_i$, $i = 0, 1, 2, \ldots, 13$. Hence, for every time step $t_i$, exactly one event is added to the network. Edges are also added to the network at the same time step as women are attending the events. Every edge that is added to the network at $t_i$ must be connected to the event with time stamp $t_i$.

Any 6-cycle in a time dependent bipartite network is formed by closing an existing 4-path by connecting a new passive node to the two active end nodes of the 4-path. Figure 5 shows all the possibilities by which the distinct 6-cycles can be formed. We call the 4-path in Fig. 5a an unconnected 4-path, the 4-path in Fig. 5b a connected 4-path and the 4-path in Fig. 5c a completely connected 4-path. Assume there exists an unconnected 4-path $p = \{v_0, w_0, v_1, w_1, v_2\}$ at time $t_i$. At time $t_{i+1}$ one passive node $w_2$ is added to the network which could close $p$ by connecting to nodes $v_0$ and $v_2$ to form an unconnected 6-cycle. A second scenario is possible; the newly added passive node not only connects to nodes $v_0$ and $v_2$, but also to node $v_1$, forming a sparsely connected 6-cycle (see Fig. 5a). Similarly a connected 4-path can form a sparsely
Figure 5: All possibilities by which the different structured 6-cycles can be formed in a time dependent network. An unconnected 6-cycle and a completely connected 6-cycle can each only originate from one, distinct 4-path. A sparsely connected and a highly connected 6-cycle each have two origins.

connected or a highly connected 6-cycle (see Fig. 5b). Finally, a fully connected 4-path can form a highly connected or a completely connected 6-cycle (see Fig. 5c).

Using the origins of 6-cycles, we define equations (3) - (6) to measure the four different clustering coefficients $cc(k)$ in a time dependent bipartite network.

The unconnected clustering coefficient:

$$cc(0) = \frac{\lambda^*(0)}{\lambda(0)}$$

where $\lambda^*(0)$ is the number of closed 4-paths that form an unconnected 6-cycle and $\lambda(0)$ is the total number of unconnected 4-paths. The unconnected clustering coefficient $cc(0)$ measures the proportion of unconnected 4-paths that are closed and form an unconnected 6-cycle.

The sparsely connected clustering coefficient:

$$cc(1) = \frac{\lambda^*(1)}{\lambda(0) + \lambda(1)}$$

where $\lambda^*(1)$ is the number of closed 4-paths that form a sparsely connected 6-cycle and $\lambda(1)$ is the total number of connected 4-paths. The sparsely connected clustering coefficient $cc(1)$ measures the proportion of 4-paths that are closed and form a sparsely connected 6-cycle.
The highly connected clustering coefficient:

\[ cc_{(2)} = \frac{\lambda^*_2}{\lambda_1 + \lambda_2}, \]  

(5)

where \( \lambda^*_2 \) is the number of closed 4-paths that form a highly connected 6-cycle and \( \lambda_2 \) is the total number of completely connected 4-paths. The highly connected clustering coefficient \( cc_{(2)} \) measures the proportion of 4-paths that are closed and form a highly connected 6-cycle.

The completely connected clustering coefficient:

\[ cc_{(3)} = \frac{\lambda^*_3}{\lambda_2}, \]  

(6)

where \( \lambda^*_3 \) is the number of closed 4-paths that form a completely connected 6-cycle. The clustering coefficient \( cc_{(3)} \) measures the proportion of 4-paths that are closed and form a completely connected 6-cycle.

The local clustering coefficients \( cc_{(i,k)} \) of a node \( v_i \) can be measured in a similar way. For example, the local clustering coefficient \( cc_{(i,0)} \) of the node \( v_i \) is measured by dividing the number of closed 4-paths that are centred at \( v_i \) and form an unconnected 6-cycle by the number of all unconnected 4-paths that are centred at \( v_i \).

This section has shown why it is important to distinguish between different types of 6-cycles which are determined by the number of additional edges. We have introduced four different clustering coefficients, one for each type of 6-cycle, and shown that it may only be applied to time dependent bipartite networks.

4 Results and Discussion

We now apply the four clustering coefficients (Eq. (3) - Eq. (6)) that were defined in Section 3 to different networks from distinct areas and analyse the obtained results. We show that the nodes who drive the clustering behaviour can be identified by comparing the local clustering coefficients to the global clustering coefficients of the network.
4.1 The Southern Women Network

The first network analysed is the southern women network \[12\] that was given as an example of a time dependent network in section 3. Table 1 shows the clustering coefficients of the southern women network and the coefficients of 50 randomly generated bipartite networks of the same size and density and their 95% confidence intervals. The random networks were generated using the R package igraph \([11]\), such that the generated networks do not have exactly the same degree sequence as the southern women network. We knowingly chose this method as the events in the southern women network do not have a restriction on how many people are allowed to attend. Similarly the women can choose how many events they would like to visit.

|                  | Southern Women Network | random network |
|------------------|------------------------|----------------|
| cc\(_{(0)}\)     | 0.4446                 | 0.6226 [0.609, 0.6362] |
| cc\(_{(1)}\)     | 0.6532                 | 0.5492 [0.5365, 0.5619] |
| cc\(_{(2)}\)     | 0.5984                 | 0.4228 [0.4049, 0.4407] |
| cc\(_{(3)}\)     | 0.5604                 | 0.3604 [0.3316, 0.3892] |

Table 1: The four clustering coefficients of the southern women network and the average clustering coefficients of 50 randomly generated networks of the same size and their 95% confidence intervals.

We found that none of the clustering coefficients lie within the 95% confidence intervals and hence, none of the values are as expected in a random network. The coefficient cc\(_{(0)}\) lies below the lower bound of the confidence interval whereas cc\(_{(1)}\), cc\(_{(2)}\) and cc\(_{(3)}\) lie above the interval.

In a random network of the same size and density as the southern women network, cc\(_{(0)}\) has the highest value with cc\(_{(0)}\) = 0.6226. In the southern women network cc\(_{(0)}\) has the lowest value with cc\(_{(0)}\) = 0.4446. Hence, a greater proportion of 4-paths are closed to form an unconnected 6-cycle in random networks than in the southern women network. Since the three other clustering coefficients lie above the 95% confidence interval, it seems that a group of three women tend to cluster if they are already connected to each other by at least one previous event. This makes perfect sense, as the southern women network is a social network and one would assume that any of the 18 women would rather attend an event with friends than by herself.

The local clustering coefficients of the southern women network are displayed in Table 2. By comparing the local clustering coefficients of each woman to the global clustering coefficients of the whole network, we can calculate a score that indicates the extent to which a woman is
Table 2: This table shows the local clustering coefficients of the 18 women and their driving scores. The little arrows at the right of each cell indicate if the value lies above or below the confidence interval of the corresponding global clustering coefficient. An equal sign indicates that the value lies within the confidence interval. The 7 women that were identified to drive the clustering behaviour of the whole network are printed in bold.

driving the clustering behaviour of the whole network. There are two cases:

1. \( cc_{(k)} < CI_{(k)} \) or

2. \( cc_{(k)} \geq CI_{(k)} \),

where \( CI_{(k)} \) is the mid point of the confidence interval that is calculated for \( cc_{(k)} \). To be able to compare the scores of different nodes, we first calculate a global driving score, denoted \( ds_{(global)} \in [0,1] \), for the network by measuring how far each of the global clustering coefficients \( cc_{(k)} \) is from the mid point and then taking the average over all four scores, see Eq. 7. The greater the difference between the global clustering coefficient and the mid point of the confidence interval, the higher the global driving score.
\[ ds_{\text{global}} = \frac{1}{4} \sum_{k=0}^{3} g(k), \quad (7) \]

where

\[ g(k) = \begin{cases} \frac{|CI_{(k)} - cc_{(k)}|}{CI_{(k)}} & \text{if } cc_{(k)} < CI_{(k)} \\ \frac{|CI_{(k)} - cc_{(k)}|}{1 - CI_{(k)}} & \text{if } cc_{(k)} \geq CI_{(k)} \end{cases}. \quad (8) \]

A woman with a local clustering coefficient \( cc_{(i,k)} \) that is close to the global clustering coefficient \( cc_{(k)} \) of the average random network, behaves as expected and hence does not contribute to a clustering behaviour that is different to a random network. If the global clustering coefficient \( cc_{(k)} \) lies below the mid point of the confidence interval (\( cc_{(k)} < CI_{(k)} \)), there are again two cases:

1. \( cc_{(i,k)} < CI_{(k)} \) or
2. \( cc_{(i,k)} \geq CI_{(k)} \).

If the local clustering coefficient \( cc_{(i,k)} \) of node \( i \) lies below the mid point of the confidence interval, then node \( i \) contributes to the global clustering behaviour of the whole network and we assign a score between 0 and 1 to node \( i \), depending on the difference between the local clustering coefficient and the mid point of the confidence interval. On the contrary, if \( cc_{(i,k)} \) lies above the mid point of the confidence interval, then node \( i \) drives against the clustering behaviour and we assign a score between 0 and -1. Similarly, if the global clustering coefficient \( cc_{(k)} \) lies above the mid point of the confidence interval, there are two cases.

The driving score, \( ds_{(i)} \), of node \( i \) is given by the following equation:

\[ ds_{(i)} = \frac{1}{4} \sum_{k=0}^{3} f(k), \quad (9) \]

where
\[ f(k) = \begin{cases} 
|CI_k - cc_{i,k}| & \text{if } cc_{(k)} < CI_{(k)} > cc_{(i,k)} \\
CI_{(k)} & \text{if } cc_{(k)} < CI_{(k)} \leq cc_{(i,k)} \\
1 - CI_{(k)} & \text{if } cc_{(k)} \geq CI_{(k)} \leq cc_{(i,k)} \\
|CI_k - cc_{i,k}| & \text{if } cc_{(k)} \geq CI_{(k)} > cc_{(i,k)} 
\end{cases} \quad (10) \]

For the southern women network \( ds_{(global)} = 0.2834 \). The driving scores of the women (see Table 2), reveal that women 1, 2, 3, 4, 7, 9 and 13 drive the clustering behaviour of the whole network as they lie above the global driving score and hence drive the clustering behaviour of the network the most. All nodes with a negative driving score drive against the global driving behaviour of the whole network. The identification of influential nodes is not possible using previous definitions of the bipartite clustering coefficient.

| Southern Women Network | random network |
|------------------------|----------------|
| \( cc_{(0)} \)       | 0.3578         | 0.7236 [0.7049, 0.7422] |
| \( cc_{(1)} \)       | 0.597          | 0.6402 [0.6283, 0.6521] |
| \( cc_{(2)} \)       | 0.8556         | 0.5113 [0.4915, 0.531] |
| \( cc_{(3)} \)       | 0.7903         | 0.4388 [0.4017, 0.4758] |

Table 3: The four clustering coefficients of the southern women network with respect to the passive node set of events and the average clustering coefficients of 50 randomly generated networks of the same size and their 95% confidence interval.

The first analysis of the southern women dataset was carried out by Davis, Gardner and Gardner [12] in the form of interviews, with the aim to categorise the 18 women into groups. They found two different groups that were further divided into core, primary and secondary members. Figure 6 shows the southern women network with the two groups identified in [12]. Our analysis found all the core nodes of the first group to be influential as well as one core node of the second group. Interestingly, our results show that women 7 and 9 should also be considered as important. Both attended only four events, however, these events were also attended by members from both groups. This observation indicates that women 7 and 9 may have been an important connection between the two groups. Davis, Gardner and Gardner also found that woman 9 had some affiliation with both groups. We further noticed that the four clustering coefficients of the two subnetworks corresponding to the two cores of each group equal 1. When a primary or secondary member of a group is added to the respective subnetwork, at least one clustering coefficient has a value less than 1. This confirms the role of women 1, 2, 3 and 4 as
core members in group 1 and women 13, 14 and 15 in group 2. Further investigation is needed in order to detect whole communities by applying the clustering coefficient but is out of the scope of this paper.

Table 4: This table shows the local clustering coefficients of the 14 events and their driving scores. The little arrows at the right of each cell indicate if the value lies above or below the confidence interval of the corresponding global clustering coefficient. An equal sign indicates that the value lies within the confidence interval. The events that were identified to drive the clustering behaviour of the whole network are printed in bold. The global driving score with respect to the events equals 0.476.

| event \(i\) | \(cc_{(i,0)}\)  | \(cc_{(i,1)}\)  | \(cc_{(i,2)}\)  | \(cc_{(i,3)}\)  | \(ds_{(i)}\)  |
|-----------|----------------|----------------|----------------|----------------|----------------|
| 1         | 1 ↑ 0.9556 ↑ 0.7714 ↑ 0.6 ↑ -0.2536 |
| 2         | 0.8 ↑ 0.9574 ↑ 0.8571 ↑ 0.5143 ↑ -0.074 |
| 3         | 0.3043 ↓ 0.7113 ↑ 0.9727 ↑ 0.8824 ↑ 0.5584 |
| 4         | 0.9 ↑ 0.9529 ↑ 0.8803 ↑ 0.6427 ↑ -0.0838 |
| 5         | 0.2545 ↓ 0.7952 ↑ 0.9895 ↑ 0.9029 ↑ 0.5364 |
| 6         | 0.3421 ↓ 0.5482 ↓ 0.8913 ↑ 0.8791 ↑ 0.5874 |
| 7         | 0.3195 ↓ 0.6965 ↑ 0.8165 ↑ 0.7051 ↑ 0.3929 |
| 8         | 0.38 ↓ 0.5918 ↓ 0.9429 ↑ 0.8672 ↑ 0.5776 |
| 9         | 0.3062 ↓ 0.6823 ↑ 0.7968 ↑ 0.6923 ↑ 0.3907 |
| 10        | 0.48 ↓ 0.7023 ↑ 0.7891 ↑ 0.8049 ↑ 0.3704 |
| 11        | 1 ↑ 0.7949 ↑ 0.1 ↓ 0 ↓ -0.8086 |
| 12        | 0.3889 ↓ 0.7348 ↑ 0.8187 ↑ 0.875 ↑ 0.4303 |
| 13        | 1 ↑ 0.6098 ↓ 0.5323 ↑ 0.6923 ↑ -0.0977 |
| 14        | 1 ↑ 0.6098 ↓ 0.5323 ↑ 0.6923 ↑ -0.0977 |

So far we have focused our analysis only on the 18 women. We now repeat the analysis for the passive node set that represents the 14 events. Table 3 shows the clustering coefficients of the southern women network with respect to the passive node set.

Again, none of the four clustering coefficients lie within the 95% confidence interval of the randomly generated networks. The coefficients \(cc_{(0)}\) and \(cc_{(1)}\) lie below the lower bound of the respective confidence interval whereas the \(cc_{(2)}\) and \(cc_{(3)}\) lie above the interval.

Comparing the local clustering coefficients of the 14 events, displayed in Table 4, to the four global clustering coefficients, we find that events 3, 5, 6 and 8 drive the clustering behaviour of the network. Figure 6 shows that these four events had a high attendance and events 5, 6 and 8 were attended by women from both groups and again, it seems that these events are influential because they connect the members of the two groups to each other.
Figure 6: This figure shows the southern women network. The two groups and their core, primary and secondary members that were identified by Davis, Gardner and Gardner [12] are labeled. The darker shaded nodes are driving the clustering behaviour of the network, identified by the proposed analysis.

4.2 The Noordin Top Terrorist Network

We now investigate a subset of the Noordin Top Terrorist network [14]. This particular subset models the attendance of 26 members of the terrorist network at 20 different meetings. The subset contains a total of 64 connections between members and meetings. Table 5 and Table 6 show the clustering coefficients of the members of the terrorist ring and the meetings respectively. The random networks were generated in the same manner as described in Subsection 4.1.

|          | Noordin Top Terrorist Network | random network                   |
|----------|-------------------------------|----------------------------------|
| $cc_{(0)}$ | 0.0303                        | 0.1913 [0.1767, 0.2059]          |
| $cc_{(1)}$ | 0.1108                        | 0.0564 [0.0469, 0.066]           |
| $cc_{(2)}$ | 0.2                           | 0.0273 [0.0139, 0.0407]          |
| $cc_{(3)}$ | 0                             | 0 [0,0.0592]                     |

Table 5: The four clustering coefficients of the terrorist network and the average clustering coefficients of 50 randomly generated networks of the same size and their 95% confidence interval with respect to the members of the terrorist ring.

For both node sets, the members and the meetings, the clustering coefficient $cc_{(0)}$ lies below the
Table 6: The four clustering coefficients of the terrorist network and the average clustering coefficients of 50 randomly generated networks of the same size and their 95% confidence interval with respect to the meetings.

|                | Noordin Top Terrorist Network | random network |
|----------------|-------------------------------|----------------|
| $\text{cc}(0)$ | 0.0656                        | 0.2497 [0.2314, 0.2681] |
| $\text{cc}(1)$ | 0.2153                        | 0.089 [0.0746, 0.1035] |
| $\text{cc}(2)$ | 0.0879                        | 0.0382 [0.0232, 0.0533] |
| $\text{cc}(3)$ | 0                             | 0.01 [0.0.0296] |

Due to the low density of the network, all clustering coefficients are low. In the terrorist network, the proportion of 4-paths that are closed and form a highly connected 6-cycle is much higher than in a random network. As in the southern women network, it seems that three members of the terrorist ring would cluster if they were already connected through at least one previous meeting. Since this network is not a social network, we cannot explain this clustering behaviour by its underlying social structure. We can however look more closely at the local clustering coefficients to determine which nodes are driving this behaviour.

The local clustering coefficients of the members of the terrorist network are shown in Table 7. The driving scores clearly show the influential nodes that are driving the clustering behaviour of the network are the members 5, 8 and 18. These correspond to Ahmad Rofiq Ridho, Azhari Husin and Noordin Mohammed Top respectively. Noordin Mohammed Top and Azhari Husin worked together to plan the terrorist attacks, with Noordin Mohammed Top financing the attacks and Azhari Husin in charge of building the bombs. Ahmad Rofiq Ridho was a communicator between the members. Member 19 also received a very high driving score and was a communicator like Ahmad Rofiq Ridho. Table 7 shows that for most nodes in the network the clustering coefficients are undefined and hence most clusters contain the driving nodes. The driving scores of the passive node set could not reveal any meeting that drives the clustering behaviour. This may be explained as follows: the members of the terrorist ring may not have been able to choose which meeting to attend. Instead they were probably told by the organisers. Another explanation may be that all events were equally important. Unfortunately, we do not
have enough information about the meetings or the terrorist ring to back up our claims.

5 Conclusion and Future Work

Many real world networks are bipartite. However, not every network measure can be directly applied to this type of network [19]. In order to analyse bipartite networks, one can either project the network to a one-mode network or redefine those networks measure that are not suitable for the analysis of bipartite networks. Since projection leads to loss of information [9, 29, 34] network measures have been redefined in terms of bipartite networks.

A measure that has received a great amount of interest is the clustering coefficient. However, only one of the existing methods of calculating the bipartite clustering coefficient, proposed in [24], measures closure between three nodes of the same type. This method defines the clustering coefficient in terms of 6-cycles and 4-paths.

This paper showed that it is important to distinguish between different types of 6-cycles that are identified by the number of additional edges that may connect nodes within the cycle. Ignoring additional edges can result in an over-count of 6-cycles. We showed that the formation of the different types of 6-cycles depends on the network. For instance, in a network where passive nodes may connect at any point in time, a sparsely connected 6-cycle can originate from an unconnected 6-cycle. This is not possible in a time dependent network where any 6-cycle can only originate from a 4-path. We defined four clustering coefficients that correspond to the different types of 6-cycles.

Applying the four clustering coefficients to real world networks gives valuable insight into how clusters are structured and why they form. The driving scores of the nodes of a network revealed the nodes that are driving the clustering behaviour and have influence on the network structure. Previous analyses supports the results we obtained in Section [3]
Table 7: This table shows the local clustering coefficients of the 26 members of the Noordin terrorist network. The little arrows at the right of each cell indicate if the value lies above or below the confidence interval of the corresponding global clustering coefficient. An equal sign indicates that the value lies within the confidence interval. The four members that were identified to drive the clustering behaviour of the whole network are printed in bold. The global driving score with respect to the members equals 0.2692.
Considering time stamps of a network, if they are available, could give further insight into the network of interest and make a more dynamic analysis of the network possible. The clustering coefficients that were introduced in this paper can only be applied to time dependent bipartite networks. Further work needs to be done on other types of networks in which 6-cycles are formed differently. We will also focus future work on finding whole communities in bipartite networks by applying the clustering coefficients proposed in this paper.

### Table 8

This table shows the local clustering coefficients of the 20 meetings of the Noordin terrorist network and their driving scores. The little arrows at the right of each cell indicate if the value lies above or below the confidence interval of the corresponding global clustering coefficient. An equal sign indicates that the value lies within the confidence interval. The global driving score with respect to the meetings equals 0.4819.

| meeting | $cc_{(c,0)}$ | $cc_{(c,1)}$ | $cc_{(c,2)}$ | $cc_{(c,3)}$ | $ds_{(i)}$ |
|---------|-------------|-------------|-------------|-------------|------------|
| 1       | 0.0769 ↓    | 0.3077 ↑    | 0 ↓         | 0 =         | 0.233      |
| 2       | 0.1667 ↓    | 0.1333 ↑    | 0 ↓         | n/a         | -0.2063    |
| 3       | n/a         | n/a         | n/a         | n/a         | n/a        |
| 4       | 0.2 ↓       | 0.2 ↑       | n/a         | n/a         | 0.1604     |
| 5       | 0.24 =      | 0.0968 =    | 0 ↓         | n/a         | -0.3175    |
| 6       | 0 ↓         | 0.1333 ↑    | 0 ↓         | n/a         | 0.0162     |
| 7       | n/a         | n/a         | n/a         | n/a         | n/a        |
| 8       | 0.5 ↑       | 0.2222 ↑    | 0 ↓         | n/a         | -0.3958    |
| 9       | 0.3333 ↑    | 0 ↓         | n/a         | n/a         | -0.5557    |
| 10      | 0.2222 =    | 0 ↓         | n/a         | n/a         | -0.4449    |
| 11      | 0.2308 ↓    | 0.3077 ↑    | 0 ↓         | 0 =         | 0.0789     |
| 12      | 0.2414 =    | 0.0571 ↓    | 0 ↓         | n/a         | -0.4417    |
| 13      | 0.167 ↓     | 0.1579 ↑    | 0 ↓         | n/a         | -0.1977    |
| 14      | 0 ↓         | 0.6364 ↑    | 0 ↓         | n/a         | 0.2003     |
| 15      | 0.1795 ↓    | 0.1481 ↑    | 0 ↓         | 0 =         | 0.0865     |
| 16      | n/a         | n/a         | n/a         | n/a         | n/a        |
| 17      | 0.4 ↑       | 0.4 ↑       | n/a         | n/a         | 0.0705     |
| 18      | 0.6667 ↑    | 0 ↓         | n/a         | n/a         | -0.7779    |
| 19      | 0.15 ↓      | 0 ↓         | n/a         | n/a         | -0.3004    |
| 20      | 0.1639 ↓    | 0.1852 ↑    | 0 ↓         | n/a         | -0.1836    |
References

[1] Asratian, AS, Denley, TM & Häggkvist, R, 1998, Bipartite graphs and their applications, 1st edn., Cambridge University Press, Cambridge.

[2] Barber, MJ, 2007, ‘Modularity and community detection in bipartite networks’, Physical Review E, vol. 76, no. 6, pp. 066102.

[3] Barthélemy, M, 2011, ‘Spatial networks’, Physics Reports, vol. 499, no. 1, pp. 1–101.

[4] BBC, 2014, ‘Obituary: Noordin Mohamed Top’, online, URL http://news.bbc.co.uk/2/hi/asia-pacific/4302368.stm

[5] Bonacich, P, 1978, ‘Using boolean algebra to analyze overlapping memberships’, Sociological Methodology, (pp. 101–115).

[6] Borgatti, SP, 2012, ‘Social network analysis, two-mode concepts in’, in ‘Computational Complexity’, (pp. 2912–2924), Springer.

[7] Borgatti, SP, Everett, MG & Johnson, JC, 2013, Analyzing social networks, 1st edn., SAGE, London.

[8] Borgatti, SP & Halgin, DS, 2011, ‘Analyzing affiliation networks’, in ‘The Sage handbook of social network analysis’, (pp. 417–433), Sage Publications, Thousand Oaks.

[9] Conaldi, G, Lomi, A & Tonellato, M, 2012, ‘Dynamic models of affiliation and the network structure of problem solving in an open source software project’, Organizational Research Methods, vol. 15, no. 3, pp. 385–412.

[10] Costa, LdF, Rodrigues, FA, Travieso, G & Villas Boas, P, 2007, ‘Characterization of complex networks: A survey of measurements’, Advances in Physics, vol. 56, no. 1, pp. 167–242.

[11] Csardi, G & Nepusz, T, 2006, ‘The igraph software package for complex network research’, InterJournal, Complex Systems, vol. 1695, no. 5, p. 1695, URL http://igraph.sf.net

[12] Davis, A, Gardner, BB & Gardner, MR, 1941, Deep South: A social anthropological study of caste and class, 1st edn., University of Chicago Press, Chicago.

[13] Diestel, R, 2005, Graph theory, 3rd edn., Springer, New York.
[14] Everton, SF, 2012, *Disrupting dark networks*, 1st edn., Cambridge University Press, Cambridge.

[15] Fortuna, MA, Popa-Lisceanu, AG, Ibáñez, C & Bascompte, J, 2009, ‘The roosting spatial network of a bird-predator bat’, *Ecology*, vol. 90, no. 4, pp. 934–944.

[16] Freeman, LC, 2003, ‘Finding social groups: A meta-analysis of the southern women data’, in ‘Dynamic social network modeling and analysis’, (pp. 39–97), The National Academics Press, Washington.

[17] Holme, P, Liljeros, F, Edling, CR & Kim, BJ, 2003, ‘Network bipartivity’, *Physical Review E*, vol. 68, no. 5, pp. 056107.

[18] Kitsak, M, Gallos, LK, Havlin, S, Liljeros, F, Muchnik, L, Stanley, HE & Makse, HA, 2010, ‘Identification of influential spreaders in complex networks’, *Nature Physics*, vol. 6, no. 11, pp. 888–893.

[19] Latapy, M, Magnien, C & Vecchio, ND, 2008, ‘Basic notions for the analysis of large two-mode networks’, *Social Networks*, vol. 30, no. 1, pp. 31–48.

[20] Lind, PG, Gonzalez, MC & Herrmann, HJ, 2005, ‘Cycles and clustering in bipartite networks’, *Physical Review E*, vol. 72, no. 5, pp. 056127.

[21] Newman, MEJ, 2001, ‘The structure of scientific collaboration networks’, *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409.

[22] Newman, MEJ, 2006, ‘Finding community structure in networks using the eigenvectors of matrices’, *Physical Review E*, vol. 74, no. 3, pp. 036104.

[23] Newman, MEJ, 2010, *Networks, An Introduction*, 1st edn., Oxford University Press, Oxford.

[24] Opsahl, T, 2013, ‘Triadic closure in two-mode networks: Redefining the global and local clustering coefficients’, *Social Networks*, vol. 34, no. 2, pp. 159–167.

[25] Piepenbrink, A & Gaur, AS, 2013, ‘Methodological advances in the analysis of bipartite networks: An illustration using board interlocks in indian firms’, *Organizational Research Methods*, vol. 16, no. 3, pp. 474–496.
[26] Robins, G & Alexander, M, 2004, ‘Small worlds among interlocking directors: Network structure and distance in bipartite graphs’, *Computational & Mathematical Organization Theory*, vol. 10, no. 1, pp. 69–94.

[27] Rubinov, M & Sporns, O, 2010, ‘Complex network measures of brain connectivity: uses and interpretations’, *Neuroimage*, vol. 52, no. 3, pp. 1059–1069.

[28] Tucker, A, 1943, *Applied combinatorics*, 6th edn., John Wiley & Sons, Inc., New York.

[29] Vogt, I & Mestres, J, 2010, ‘Drug-target networks’, *Molecular Informatics*, vol. 29, no. 1-2, pp. 10 – 14.

[30] Wasserman, S & Faust, K, 1994, *Social network analysis: Methods and applications*, 1st edn., Cambridge University Press, Cambridge.

[31] Zhan, W, Zhang, Z, Guan, J & Zhou, S, 2011, ‘Evolutionary method for finding communities in bipartite networks’, *Physical Review E*, vol. 83, no. 6, pp. 066120.

[32] Zhang, P, Wang, J, Li, X, Li, M, Di, Z & Fan, Y, 2008, ‘Clustering coefficient and community structure of bipartite networks’, *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 27, pp. 6869–6875.

[33] Zhang, X, Zhu, J, Wang, Q & Zhao, H, 2013, ‘Identifying influential nodes in complex networks with community structure’, *Knowledge-Based Systems*, vol. 42, pp. 74–84.

[34] Zhou, T, Ren, J, Medo, M & Zhang, YC, 2007, ‘Bipartite network projection and personal recommendation’, *Physical Review E*, vol. 76, no. 4, pp. 046115.