Alzahrani and Horadam *Applied Network Science* (2019) 4:21
https://doi.org/10.1007/s41109-019-0123-6

Applied Network Science

RESEARCH Open Access

Finding maximal bicliques in bipartite networks using node similarity

Taher Alzahrani* and Kathy Horadam

*Correspondence:
t.alzahrani@mof.gov.sa
Mathematical Sciences, RMIT University, Melbourne, Australia

Abstract

In real world complex networks, communities are usually both overlapping and hierarchical. A very important class of complex networks is the bipartite networks. Maximal bicliques are the strongest possible structural communities within them. Here we consider overlapping communities in bipartite networks and propose a method that detects an order-limited number of overlapping maximal bicliques covering the network. We formalise a measure of relative community strength by which communities can be categorised, compared and ranked. There are very few real bipartite datasets for which any external ground truth about overlapping communities is known. Here we test three such datasets. We categorise and rank the maximal biclique communities found by our algorithm according to our measure of strength. Deeper analysis of these bicliques shows they accord with ground truth and give useful additional insight. Based on this we suggest our algorithm can find true communities at the first level of a hierarchy. We add a heuristic merging stage to the maximal biclique algorithm to produce a second level hierarchy with fewer communities and obtain positive results when compared with other overlapping community detection algorithms for bipartite networks.

Keywords: Bipartite network, Overlapping community detection, Maximal biclique, Community strength, Node similarity

Introduction

The main contribution of this paper is an algorithm combining three concepts (node similarity, maximal bicliques and cliques) that can improve community detection in bipartite networks. The algorithm we introduce, **MaxBic**, produces overlapping maximal bicliques, covers the network and forms the base level of a community hierarchy. Structurally, these bicliques are as tightly connected internally as is possible in the network. We measure how relatively strong or weak they are as communities within the network, according to 5 categories of community strength, formalised here. MaxBic is a deterministic algorithm and requires no predefined parameters such as the number of communities, maximum number of community memberships, or allowed proportion of overlap, as initial input. For a network with n nodes, it produces no more than n maximal bicliques. We show its time complexity is at worst $O\left(n^3\right)$, irrespective of whether the network is dense or sparse.

A network G is *bipartite* if its nodes can be partitioned into two sets P (the primary set) and S (the secondary set) such that edges can only occur between nodes in different sets. The most tightly connected node sets which can be found in G are complete bipartite

graphs, or *bicliques*, on subsets of a nodes in P and b nodes in S, denoted $K_{a,b}$. A biclique is *maximal* in G if it is not a proper subgraph of another biclique in G.

At present there is no commonly accepted standard to evaluate the efficiency and accuracy of overlapping community detection algorithms for bipartite networks. Fundamentally then, the meaningfulness of overlapping communities can only be assessed using external metadata analysis or ground truth where it exists. There are very few such bipartite networks known to us, and as far as we know, this is the first study based on having some external ground truth (for $P \vee S$ or P or S) for three networks. These are the (small) benchmark Southern Women social network, the Noordin Top terrorist network and the NSW crime area network. Deeper analysis of the maximal bicliques MaxBic detects in $P \vee S$ for these networks shows they determine groups that, while smaller and more numerous than the ground truth communities, are meaningful and bring new insights. Based on these results, we make the assumption that MaxBic's bicliques do represent real communities at the base level. We compare them with those found by other algorithms for the Southern Women network. To reduce the number of communities detected, we apply a second stage merging algorithm based on Jaccard similarity. Then we compare the performance of our two-stage algorithm **MaxBicR** on the three real networks with that of other algorithms in the literature. We show it has improved performance in community detection.

The paper is organized as follows. The rest of this section discusses related work. In "Methods" section the "Community strength" subsection contains our definition of community strength, "MaxBic: a new maximal biclique finding algorithm" subsection formally describes the methodology of our biclique-finding algorithm MaxBic, and "Computational complexity" subsection provides an overview of its computational complexity. We introduce our heuristic second-stage algorithm to reduce redundancy of communities in "Reducing redundancy and revealing hierarchy–MaxBicR (J)" subsection. In "Results" section we apply MaxBic to the three real networks, evaluate the communities it finds against the ground truth, and uncover new insights. We compare its performance with other algorithms on the benchmark Southern Women network. Then we compare the performance of our two-stage algorithm with others on the three networks, taking MaxBic communities as base level ground truth. Finally, we summarise and discuss future work in "Conclusion and future work" subsection.

Overlapping communities algorithms

The problem of detecting communities in networks has a long and rich history, particularly for social networks, where it began in the 1940s (Luce and Perry 1949). For instance, one widely applied approach for complex networks detects weak links and cuts them to separate the communities, using e.g. modularity (Newman 2006; Newman and Girvan 2004) or betweenness centrality (Girvan and Newman 2002). This does not take into account the possibility that a node might belong to more than one community, so is unrealistic for many real networks. For example, the original Label Propagation Algorithms (LPAs), introduced by Raghavan et al. (2007) and followed by Leung et al. (2009), assign only one membership to every node in the network.

Work on detection of overlapping communities, or fuzzy clusters, goes back at least to the 1970s, e.g. Ruspini (1970). The sociological concept of structural cohesion was formalised in Moody and White (2003) in terms of connectivity: the structural cohesion

of a network equals the minimum number of nodes whose removal would disconnect the network. A clique has maximum structural cohesion, as every node must be removed to disconnect it. In a bipartite network, no cliques exist and a biclique has maximum structural cohesion: every node in the smaller set (P or S) must be removed to disconnect it.

There has been much more emphasis on detecting overlapping communities in general (unipartite) networks than specifically in those with bipartite structure. Surveys appear in Fortunato (2010), Ch.11 and Xie et al. (2013), with a loose breakdown into link-clustering and node-clustering techniques.

The link clustering technique of Ahn et al. (2010) creates a dendrogram from a network then cuts this dendrogram at various thresholds according to a Partition Density quality function. Each link in the network is a leaf of the dendrogram whose branches represent link communities. This method involves similarity between edges rather than nodes and has been implemented as *linkcomm* (Kalinka and Tomancak 2011). It can be applied to bipartite networks but has not been specifically designed for them.

In Palla et al. (2005) the Clique Percolation Method (CPM) uses k-clique percolation with $k \geq 3$, and the overlap forms between communities where a node is in more than one clique. An exhaustive search for connected subgraphs of two or more cliques is conducted. A clique with $k = 2$ will not be detected so some nodes may not belong to any community. This may result in incomplete cover of a network (MaxBic communities cover all nodes of the network). In Lázár et al. (2010), a quality function M^{ov} is developed to overcome the limitation of Newman-Girvan Q-modularity (Newman and Girvan 2004) with overlapping communities. The authors compare the performance of the CFinder algorithm of Palla et al. (2005) with a version of the link-clustering algorithm of Ahn et al. (2010), using M^{ov}. Our community strength measures are simpler, but similar in approach to Lázár et al. (2010). Another technique, intrinsic Longitudinal Community Detection (iLCD) (Cazabet et al. 2010), discovers highly overlapping groups of nodes. This approach takes the dynamics of the network into consideration and is claimed to be preferable to CPM in terms of efficiency and computational time. Evans (2010) shows that detecting overlapping communities using the cliques graph results in more significant structural communities than detecting overlapping communities directly using node clustering approaches. The cliques of order k are found, then a network with these cliques as nodes is constructed. The node clustering algorithm applied to the clique graph is *Louvain* (Blondel et al. 2008). Note that if *Louvain* is run past its first phase, it may fail to find smaller cliques (Lancichinetti and Fortunato 2014). The very successful *Infomap* algorithm (Rosvall and Bergstrom 2008) has an option to identify overlapping flow communities (Esquivel and Rosvall 2011). Clique-finding algorithms are not effective on bipartite networks.

We next survey recent techniques designed for bipartite networks.

A recent study (Tarissan 2015) aims to discriminate between two metrics defined on set P, the bipartite clustering coefficient and the bipartite redundancy coefficient, which attempt to measure the amount of overlap between communities. They favour the redundancy as they find the clustering correlates with node degree in several real networks. In Xu et al. (2013) an algorithm for community detection in bipartite networks is proposed based on ant colony optimization. It is tested on the Southern Women network but determines communities in P and in S separately, not truly bipartite communities.

The LPA technique has been extended by Gregory (2010) in the Community Overlap PRopagation Algorithm (COPRA) to detect overlapping communities in unipartite and bipartite networks, by introducing a parameter v, the maximum number of community memberships per node. For sparse networks, with approximately equal numbers of nodes and edges, and with small v, COPRA is a very fast algorithm: the time complexity $O\left(v^3 n\right)$, plus $O(vnlog(v))$ per iteration, is almost linear in the network order. A heuristic adaptive LPA, BiLPA, is developed by Li et al. (2016) specifically for bipartite networks. They define a bipartite partition density function and do not need to prespecify the number of communities to be found, but do have to prespecify a threshold θ for a measure based on neighborhood overlap above which a node will be assigned to a community. (MaxBic does not require any parameters to be input.)

Probably the most investigated technique is biclique-finding and merging, in view of the maximum structural cohesiveness inherent in a biclique. In Cui and Wang (2014) the authors progressively merge minimal bicliques (of the form $K_{a,1}$: degree a node s in S has a neighbourhood of a nodes in P) when the current merged set of nodes overlaps the a nodes and the ratio of overlap number to the degree of s exceeds 0.5. After iteration, any unmerged nodes are subsequently allocated to all communities they connect to. Their results on the Southern Women network show communities in P in accord with results obtained by unipartite overlapping algorithms, but their results on other small bipartite networks are presented without analysis. They claim a time complexity of $O\left(n^2\right)$. By contrast, MaxBic generates all bicliques of the form $K_{2,b}$ (which we call basic bicliques) before merging begins.

In Lehmann et al. (2008) the authors extend the idea of a k-clique community from CPM to bipartite graphs: a $K_{a,b}$ biclique community is the union of all $K_{a,b}$ bicliques that can be reached from each other, through a series of adjacent $K_{a,b}$ bicliques. Two $K_{a,b}$ bicliques are adjacent if their overlap is at least a $K_{a-1,b-1}$ biclique. Du et al. (2008) worked with ideas related to biclique overlap, and named their algorithm "BiTector". In BiTector, initially *all* maximal bicliques are extracted in order to use them as "clustering cores". Then communities are built up by expanding and merging the clustering cores according to a closeness function based on Jaccard similarity of node sets. On sparse bipartite networks BiTector is claimed (somewhat surprisingly) to have time complexity approximately proportional to $O(n)$ for the extraction of all maximal bicliques, and overall time complexity $O\left(n^2\right)$. It would be relatively slow on a dense network. BiTector does not require any input parameters and covers the network.

The maximal biclique generation problem (MBGP), that of generating all the maximal bicliques of a network, cannot be solved in polynomial time with respect to n. It is at least as hard (Alexe et al. 2004, Lehmann et al. 2008) as the problem of finding a biclique with a maximum number of edges, the decision version of which is NP-complete (Peeters 2003). As with enumerating maximal cliques, MBGP can be solved at least exponentially in n (Viard et al. 2016). However some classes of bipartite networks have only polynomially many maximal bicliques (Alexe et al. 2004), and in some classes of bipartite networks, variants of MBGP have polynomial solutions (Makin and Uno 2004).

In contrast to BiTector, MaxBic finds at most n maximal bicliques, derived from merging node sets based on nodes with optimal similarity in a unipartite supergraph G^* of G, but it may be slower than BiTector on sparse networks.

Methods

We denote a bipartite network by $G = (P, S, E)$ where $P = \{p_1, p_2, .., p_k\}$ and $S = \{s_1, s_2, .., s_l\}$, $E \subseteq P \times S$ is the set of edges in G, $n = k + l$ is the order of G and $m = |E|$ is the size. Every bipartite network we deal with is assumed to be unweighted and undirected, so its adjacency matrix A is symmetric and can be written in block form. It is not necessary that it be connected, since each component can be separately clustered, but for simplicity here we assume that it is (i.e. there exists a path between every pair of nodes). We further assume $k \geq 2$ and $l \geq 2$, since otherwise G is a star and necessarily forms a single community. For brevity, we may represent a biclique by its node set alone.

The MaxBic algorithm is based on three concepts: node similarity; transformation of a biclique into a clique on the same node set to simplify tracking of overlap and merging; and maximal bicliques. We work with node similarity, rather than edge similarity as in Ahn et al. (2010), because in social networks nodes in the same community have similar patterns and a community of nodes has homogenous structure Barrat et al. (2008).

Node similarity

In social science, the idea of similarity between nodes is not new, with studies going back to 1971 (Lorrain and White 1971). Similarity there is defined in term of structural equivalence (Lorrain and White 1971; Van Steen 2010) where two nodes i and j are structurally equivalent if they have the same pattern of relationships with all other nodes. This implies that nodes i and j share the same neighbors for the same purpose (Leicht et al. 2006). The more similar the nodes are, the more common neighbors they have.

In a bipartite network, it is usual to infer that the more neighbors that two nodes i and j in the same set (say P) have in common in the other set (S), the higher the likelihood that they interact, and so the structural similarity of nodes i and j can be measured by their number of common neighbors. It has been proved that this count is an effective measure for structural similarity and gives accurate results (Zhou et al. 2009; Liben Nowell and Kleinberg 2007) on large-scale networks. It underpins the definition of the unipartite projections G_P and G_S of G. Moreover, it uses the fundamental topology of the network (Leicht et al. 2006). Thus we define the *similarity of nodes i and j* to be their *common neighbors index* CNI_{ij}, i.e. the size of their set of common neighbors $CNS_{ij} \equiv \Gamma(i) \cap \Gamma(j)$, where $\Gamma(i) \equiv \{x | \{i, x\} \in E\}$ is the exclusive neighborhood of node i. The similarity can be calculated either from the CNS or from the network's adjacency matrix $A = [a_{ij}]$:

$$CNI_{ij} \equiv |\Gamma(i) \cap \Gamma(j)| = \sum_x a_{ix} a_{jx} . \tag{1}$$

Community strength

Clusters of nodes can be regarded as strong or weak. Probably the simplest and most natural definition of a strong cluster is a set of nodes which form a clique, that is, the subgraph they induce is complete (Palla et al. 2005). The definition of modularity in Newman and Girvan (2004) is given for general networks and compares the number of edges within a cluster to the expected number in an equivalent network with edges placed at random, so a clique will maximise modularity for its set of nodes. Bimodularity is correspondingly defined for bipartite networks (Barber 2007), so that a biclique will maximise it. However there are less absolute ideas of community which are commonly used, and which are more appropriate for measuring the relative strength of maximal bicliques.

The definitions of strong and weak community given in Radicchi et al. (2004), which compare the number of edges outgoing from the cluster to the rest of the network, are less strict conditions than those given in Hu et al. (2008), which compare the number of edges outgoing from the cluster to each other cluster and not to all the rest of the network.

Here we order community strength into 5 categories, by comparing the definitions in Hu et al. (2008); Radicchi et al. (2004).

Definition 1 *For a particular cluster c to which node i belongs, separate the degree k_i of i into two parts: the number of edges $k_i^{in} = \sum_{j\in c} a_{ji}$ connecting node i to other nodes in c, and the number of edges $k_i^{out} = \sum_{j\notin c} a_{ij}$ connecting node i to the nodes in the rest of the network. Then c is*

1 ***strong*** *(= strong in Radicchi et al. (2004)) if*

$$k_i^{in} > k_i^{out}, \qquad \forall i \in c; \tag{2}$$

2 ***almost strong*** *(= strong in Hu et al. (2008)) if*

$$k_i^{in} \geq \max_{c'\neq c}\left\{\sum_{j\in c'} a_{ij}\right\}, \qquad \forall i \in c; \tag{3}$$

3 ***almost weak*** *(= weak in Radicchi et al. (2004)) if*

$$\sum_{i\in c} k_i^{in} > \sum_{i\in c} k_i^{out}; \tag{4}$$

4 ***weak*** *(= weak in Hu et al. (2008)) if*

$$\sum_{i\in c} k_i^{in} \geq \max_{c'\neq c}\left\{\sum_{i\in c}\sum_{j\in c'} a_{ij}\right\}, \quad \text{and} \tag{5}$$

5 ***very weak*** *if it does not belong to any strength level according to the above 4 categories.*

Clearly the definition of strong community in (2) implies the definition of almost strong in (3), which means that the definition of strong community in (2) is more restrictive than in (3). Similarly, the definition of almost weak community in (4) implies the definition of weak community in (5), which means that the definition of weak community in (5) is less restrictive than the definition in (4). Furthermore, strong implies almost weak, but almost strong does not imply almost weak. Figure 1 illustrates which definition implies which.

Based on these categories, we introduce the following measure *St* of community strength.

Definition 2 *For node i in community c let $k_i^{max-out} = \max_{c'\neq c}\sum_{j\in c'} a_{ij}$ be the maximum number of outgoing edges from i toward another community in the network. The strength $St(c)$ of c is defined as*

$$St(c) = \begin{cases} \sum_{i\in c} k_i^{in} - \sum_{i\in c} k_i^{out} & \text{if } c \text{ is strong} \\ \sum_{i\in c} k_i^{in} - \sum_{i\in c} k_i^{max-out} & \text{if } c \text{ is almost strong} \\ \sum_{i\in c} k_i^{out} - \sum_{i\in c} k_i^{in} & \text{if } c \text{ is almost weak} \\ \sum_{i\in c} k_i^{max\text{-}out} - \sum_{i\in c} k_i^{in} & \text{if } c \text{ is weak} \end{cases} \tag{6}$$

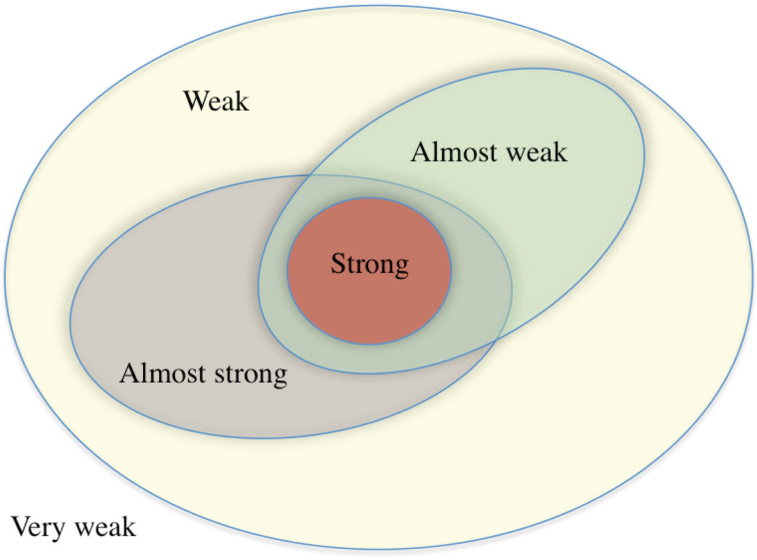

Fig. 1 Relationships between the 5 categories of community strength

The higher value for strong and almost strong communities indicates the stronger community, while the higher value for almost weak and weak communities indicates the weaker community.

MaxBic: a new maximal biclique finding algorithm

The MaxBic algorithm can be divided into seven phases, described below. Pseudocode is given in Appendix A. A worked example on a toy bipartite network is given in Fig. 2.

MaxBic finds, from each node, at most one maximal biclique containing it (others will usually be found from different nodes). First, using the idea of node similarity, it finds the set of common neighbors for every pair of nodes in P (S can equally well be chosen, see Remark 1 below). Each pair in P together with its common neighbors is the node set of a biclique we term a *basic* biclique. Second, the node set of each basic biclique is formally treated as the node set of a clique in a (unipartite) supergraph G^* of G, in order to merge node sets based on (bi)cliques rather than merging individual nodes. Conceptually this is based on the results from (Evans 2010) which show in benchmark unipartite networks that clique graphs find overlapping communities accurately while node partition methods fail. Third, the inclusive neighborhoods for each node in G^* are cross-checked for complete overlap to permit merging and enlarging, based on the idea that two nodes in $P \vee S$ are more likely to belong to the same bipartite community in G when their node similarity in G^* is optimal. The resulting bicliques in G are checked for cover and maximality. Any node from S not yet accounted for is included in any adjacent community at this point. These communities are then categorised and ordered by their strength, as described in "Community strength" subsection.

Phase (i) *Determine the common neighbors set of each pair of nodes* $p, p' \in P$. Find the exclusive neighborhood $\Gamma(p)$ for each $p \in P$, then, for every pair of nodes in p, p' in P, find $CNS_{pp'}$. The subgraph of G induced by the node set $\{p, p', \Gamma(p) \cap \Gamma(p')\}$ is a *basic biclique* $K_{2,CNI_{p,p'}}$.

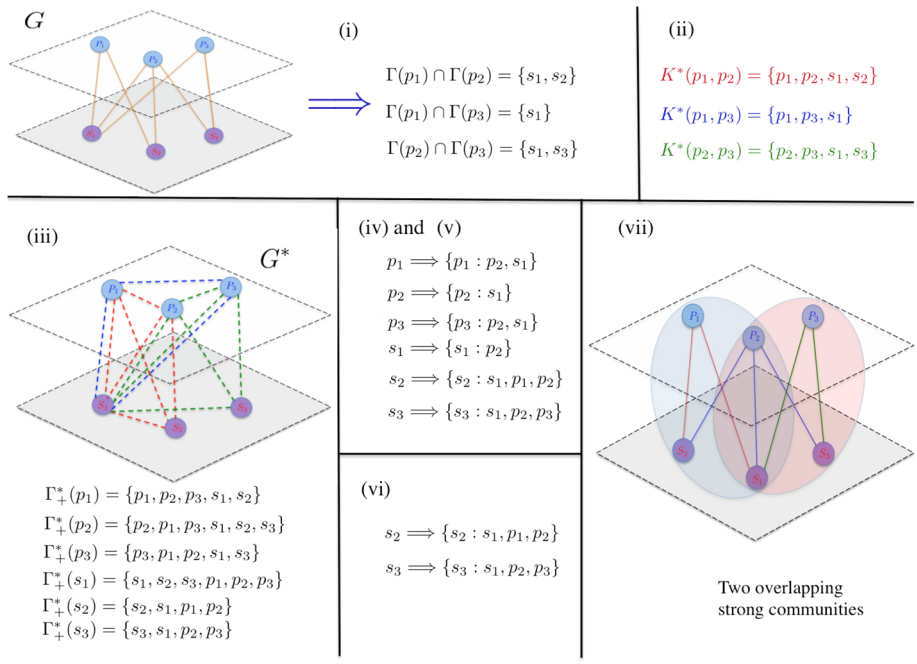

Fig. 2 Schematic illustration on toy example. In (**i**) we find basic bicliques, and then in (**ii**) we treat them as cliques. In (**iii**) we find inclusive neighborhoods for each node in cliques. In (**iv**) and (**v**) comparison between sets of nodes detects structural similarity between sets in order to merge them into larger bicliques. In (**vi**) the final two sets after merging are presented, and finally in (**vii**) the two overlapping strong communities are shown

Phase (ii) *Replace each basic biclique by a clique.* Connect each node in $K(p,p') = \{p,p',\Gamma(p)\cap\Gamma(p')\}$ to every other node in it, to form the *basic clique* $K^*(p,p')$. Each basic clique is underpinned by the structure of the projections G_P and G_S: added edge $\{p,p'\}$ in G^* would form an edge in G_P and the added edges on $CNS_{pp'}$ in G^* would form a clique in G_S, and so represents fundamental community structure in S. Formally we have a new network $G^* = (P \vee S, E^*)$ where E^* is the edge list of the basic cliques. Note G^* is no longer bipartite, though it has the same node set $P \vee S$ as G, and $E \subset E^*$.

Phase (iii) *Find the inclusive neighborhood of each node in G^* by merging basic cliques.* For each node $i \in G^*$ we find the set containing i and its neighbors in G^*:

if $i \in P$ then,

$$\Gamma^*_+(i) = \{i\} \cup \left(\bigcup_{p\in P,\ CNS_{ip}\neq\emptyset} K(i,p) \right) \tag{7}$$

and if $i \in S$, then

$$\Gamma^*_+(i) = \{i\} \cup \left(\bigcup_{p,p'\in P,\ i\in K(p,p')} K(p,p') \right). \tag{8}$$

Form a container Z of clusters, initially one for each node $i \in P \vee S$, which will be updated. Each cluster c_i in Z will consist of an element (node i) and a list, with the form $c_i = \{i : \text{list}_i\} = \{i : j_1, j_2 \ldots\}$. Initially $c_i = \{i : \emptyset\}$, so we start with as many clusters as we have nodes in the network G.

Phase (iv) *Test for total overlap of neighborhoods and form larger clusters accordingly.* Fix a *main* node $i \in P \vee S$ and run through each other node j, in order to merge clusters on nodes with high similarity. Merge j to list_i if

$$\Gamma_+^*(i) \subseteq \Gamma_+^*(j). \tag{9}$$

Node j will merge only if in G^* it is in a common clique with i, so in Z the nodes of $c_i = \{i : j_1, j_2 \dots\}$ after this phase will be from a clique in G^* and thus from a biclique in G (not necessarily maximal).

Phase (v) *Repeat* **Phase (iv)** *for each* $i \in P \vee S$. The output from this step is clusters, stored in Z. Each cluster will be either the node set of a biclique in G or of the form $\{s : \emptyset\}$ for some $s \in S$. (No cluster of the form $\{p : \emptyset\}$ for some $p \in P$ will be in Z: since G is connected, a nonempty $K(p, p')$ will exist).

Phase (vi) *Reduce the redundancy of clusters.* First, remove clusters c_i from Z that satisfy the condition $(i \cup \text{list}_i) \subseteq (j \cup \text{list}_j)$ for any c_j. This ensures that every $i \in P \vee S$ is in at least one cluster and that the biclique underlying the cluster is maximal.

Second, if $|\text{list}_i| \geq 2$ and $\text{list}_i \subset \text{list}_j$, merge element i to c_j. A cluster surviving to this point with $|list| < 2$, which has one node from P and the other from S, will stand alone, as it forms the smallest biclique, a single edge.

Finally, there might be some nodes s in S which are not merged, because they have degree 1 and aren't in any $\Gamma_+^*(i)$, $i \in P$, so $c_s = \{s : \emptyset\}$ in Z after Phase (v). They are included in any community to which their adjacent node (in P) belongs.

Phase (vii) *Categorise and order the communities in G based on their strength.* Using the definitions in "Community strength" subsection, we have five categories: strong, almost strong, almost weak, weak and very weak. Communities in the first four categories are ordered in descending order of strength *St*.

Remark 1 *If we use set S instead of P, the only difference in the edge list* E^* *will come from the smallest basic bicliques, those with 3 nodes. These can only make differences to the communities of size 1, 2 or 3 found by MaxBic, according to the following argument.*

Suppose $K(p_1, p_2) = \{p_1, p_2, s_1, s_2, \dots, s_{l'}\}$ *is a basic biclique when starting from P. Then* $E^* \setminus E$ *contains the edges* $(p_1, p_2), (s_1, s_2)$ *and if* $l' = 2$ *then starting from S the same edges arise by symmetry. If* $l' = 3$ *then also* (s_1, s_3) *and* (s_2, s_3) *are in* $E^* \setminus E$. *Starting from S, we would obtain basic bicliques* $K(s_1, s_2) = \{s_1, s_2, p_1, p_2, \dots\}, K(s_1, s_3) = \{s_1, s_3, p_1, p_2, \dots\}$ *and* $K(s_2, s_3) = \{s_2, s_3, p_1, p_2, \dots\}$, *so the 4 specified edges are in* $E^* \setminus E$. *If* $l' > 3$ *this argument generalises. In the smallest case* $l' = 1, (p_1, p_2)$ *is in* $E^* \setminus E$ *when starting from P but this edge will not arise when starting from S.*

Computational complexity

Enumeration of all maximal bicliques is at least exponential in n (Viard et al. 2016). However MaxBic does not find all maximal bicliques, but at most one maximal biclique for each node in G, and its time complexity is at most $O\left(n^2\, k\right)$ where $k = |P|$, according to the following argument.

Phase (i) Finding basic bicliques (Algorithm lines 5–12) requires computation of CNS for every pair of nodes in P so has time complexity $O\left(k^2\right)$.

Phase (ii-iii) Finding $\Gamma^*_+(i)$ requires a check of at worst each $K(p,p')$ (see Eqs. (7) and (8), Algorithm line 23) and there are n nodes to check for so this step has time complexity $O\left(n\,k^2\right)$.

Phase (iv-v) Searching for similarity between nodes using $\Gamma^*_+(i)$ (Algorithm lines 25–31) means comparing every pair of these, which can cost $O\left(n^2\right)$.

Phase (vi) The last operation compares each community with each other in order to reduce the redundancy (Algorithm lines 35–38) which takes $O\left(|Z|^2\right)$, where $|Z|$ is the number of communities, and is no more than n. So this step takes up to $O\left(n^2\right)$ also.

Putting these together and noting $k < n$ gives at worst $O\left(n^2\ k\right)$.

MaxBic uses set P to find up to n maximal bicliques. If $l \ll k$ it could be advantageous to use S instead of P, so overall we have complexity $O\left(n^2\ \min\{k,l\}\right)$.

Reducing redundancy and revealing hierarchy–MaxBicR(*J*)

At present there is no commonly accepted standard for evaluating the efficiency and accuracy of overlapping community detection algorithms for bipartite networks. We are proposing MaxBic as an algorithm for determining ground truth or metadata communities of overlapping maximal bicliques at the base level of a hierarchy. These communities are necessarily as tightly connected as is possible. To reduce redundancy at the base level and determine higher levels of a hierarchy of overlapping communities any suitable merging algorithm (such as described in "Overlapping communities algorithms" subsection) could be applied.

We have chosen to develop our own merging algorithm in order to avoid bias when comparing performance with other algorithms. We tried six different methods of merging base-level communities, all based on the *Jaccard similarity coefficient* of pairs of communities which, for two node sets c and c' is:

$$J(c,c') = \frac{|c \cap c'|}{|c \cup c'|} \tag{10}$$

When $J(c,c') = 1$ the node sets are identical and when $J(c,c') = 0$ they have no overlap.

We tested all six methods on the Southern Women and Noordin Top networks. For different threshold values of J, we plotted the number of second-level communities found against the extended normalised mutual information *NMI* (Lancichinetti et al. 2009) of those second-level communities and the base-level communities found by MaxBic. The larger the *NMI* value, the better the match between two structures. For simplicity here, we selected one of the six, which uses all nodes in a community so as not to bias towards P or S, and which gives relatively consistent performance. The full results are reported in Alzahrani (2016).

There are two stages in our selected merging algorithm. In the first stage, select a threshold $J \in [\,0,1\,]$ for the Jaccard similarity coefficient and merge (without discarding) any two node clusters output by MaxBic which have similarity coefficient at least J. In the second stage, treat the resulting clusters as super nodes and weight edges between two super nodes by their Jaccard coefficient, again thresholding on J. We obtain the incidence matrix of the cluster graph. We use the modified version of the Breadth-First Search algorithm (BFS) to traverse between super nodes and identify connected components of the cluster graph. The node set of each connected component is output as a community at the second level of the hierarchy. We term the combined algorithm (MaxBic followed by this

merging algorithm) MaxBicR(J). After empirical testing for J in steps of 0.1 we fix $J = 0.6$ in what follows.

Results

Our goal in this section is to demonstrate that the biclique communities captured by MaxBic are meaningful in real terms, and, assuming this, that MaxBicR can find overlapping communities better than some competing algorithms.

In order to evaluate our algorithm we have examined in detail three real bipartite social networks for which some external ground truth information or metadata analysis about communities in either set P, or set S, is available. The first, the Southern Women social network, is a de facto benchmark for testing community detection algorithms in bipartite networks. The second is a terrorist network and the third a crime-location network. For both the latter, only partitioning algorithms have been applied to date, so our detection of overlapping communities is new. We stress that we are not aware of any other real bipartite databases for which any external validation of overlapping communities exists prior to application of an overlapping community detection algorithm. We believe this is the first time the overlapping communities detected in three databases have been evaluated against prior information about their communities. The Southern Women database is typically the only one tested by other authors. We use the extended *NMI* to compare community structures.

In overlapping communities, nodes may belong to several communities, and so it is possible to measure the importance of a node in a bipartite network based on the number of communities to which it belongs. A node belonging to only one community is likely to be peripheral, and to many, to be core. Here we propose a simple statistical measure for determining if a node is core, peripheral or neither, based on the number of communities to which it belongs.

Definition 3 *Let $m(v)$ denote the number of communities to which node v belongs, and let μ be the mean and σ the standard deviation of the list of membership counts. Then v is a core node if*

$$m(v) > \mu + t_c\sigma \tag{11}$$

and v is a peripheral node if

$$m(v) < \mu - t_p\sigma\,. \tag{12}$$

Here t_p and t_c are parameters which can depend on metadata or ground truth. We would expect t_p to be large enough that nodes with $m(v) = 1$ are peripheral and t_c to be large enough that hubs are core. We let t_c and t_p be chosen by the researcher.

Benchmark “Southern Women" network

The small “Southern Women" network collected by Davis et al. (1941) has become a benchmark for testing community detection algorithms on bipartite networks. Its community structure is widely analyzed by social network researchers (Freeman 2003). This network has $k = 18$ women (who form set P) who attended $l = 14$ different events (set S), with $n = 32$ and $m = 89$. See Table 8 in Appendix A for its adjacency matrix. The two overlapping ground truth communities in P, identified on the basis of inteviews with the

women, are Women $1-9$ and Women $9-16$, which overlap on Woman 9 only. There is no ground truth published for overlapping communities in the network as a whole.

MaxBic detects 16 overlapping communities in the Southern Women network, see Table 1. Although structurally each of the 16 is a maximal biclique, hence as strongly connected internally as is possible, within the network there are no strong communities according to Definition 1. There are 2 almost strong communities, 3 almost weak communities, 7 weak and 4 very weak communities.

We analyse these to demonstrate that the communities found by MaxBic represent real information. We also compare our results with the results in the literature that used bipartite modularity-based algorithms and other techniques.

First, in 9 of our 16 communities all the women belong to the first ground truth community and in 6 they all belong to the second. For example, in our community 9, all 6 women belong to the first ground truth community and in our community 10 all 5 women belong to the second ground truth community. Only community 5 is split, which we suggest is because Events 8 and 9 that the 9 women coattended were the most popular events.

In terms of core and peripheral nodes in the whole network, using the formulas (11) and (12) we have $\mu \approx 3.72$ and $\sigma \approx 2.55$. Selecting $t_p = 1$ (to isolate nodes with membership 1), and $t_c = 1$ for consistency, we have $\mu + \sigma = 6.27$ and $\mu - \sigma = 1.17$. Therefore, the core nodes are those with membership 7 or more (Women 1-*Evelyn* and 3-*Theresa*; *Events 8, 9* and *5*) and peripheral nodes are those with membership 1 (Women 6-*Frances*, 7-*Eleanor*, 16-*Dorothy*, 17-*Olivia* and 18-*Flora*; *Events 1, 2* and *11*). *Evelyn* and *Theresa* were frequently identified in earlier studies (Freeman 2003) as core members of one of the ground truth communities. That was because they coattended 7 events.

We compare our results for the Southern Women network with results in the literature produced by overlapping community algorithms.

First, we consider published results where only the communities in set P (Women) are detected. The result of calculating *NMI* for these published community structures can be seen in Table 2.

Table 1 The 16 metadata communities found by MaxBic in the Southern Women network

Category	Comm. No.	*Events* ; Women	St
Almost strong	1	*3 4 5* ;1 3 4 5	10
	2	*3 4 5 7* ; 3 4 5	9
Almost weak	3	*9 10 12 13 14* ; 12 13 14	-10
	4	*3 5 6 8* ; 1 2 3 4 6	-10
	5	*8 9* ; 1 3 8 9 10 11 12 13 16	0
Weak	6	*2 3 5 6 8* ; 1 2 3	-11
	7	*5 6 7 8* ;2 3 4 7	-11
	8	*1 3 5 6 8* ;1 2 4	-10
	9	*3 5* ; 1 2 3 4 5 6	-7
	10	*10 12* ; 11 12 13 14 15	-6
	11	*8 9 10 12* ;11 12 13	-4
	12	*8 9 10 12 13 14* ; 12 13	-1
Very weak	13	*9 11* ; 14 17 18	-
	14	*7 8 9 12* ;10 13	-
	15	*5 7 8 9* ; 3 9	-
	16	*6 8 9* ; 1 3 8	-

Table 2 Comparison of ground truth with MaxBic, MaxBicR(0.6) and four other methods for detecting overlapping communities of Women (P)

Method	No. of communities	*NMI*
Ground truth	2	1
Probabilistic model (Chang and Tang 2014)	2	0.869604
COPRA (Gregory 2010)	2	0.666013
MaxBic	16	0.460668
MaxBicR(0.6)	9	0.336311
BiTector (Du et al. 2008)	4	0.301657
linkcomm (Ahn et al. 2010)	7	0.238834

The results in Table 2 for MaxBic and MaxBicR are very encouraging. The two methods with higher *NMI* have *prematched* the ground truth by presetting either the number of communities $K = 2$ (Chang and Tang 2014) or the maximum number of memberships $\nu = 2$ (Gregory 2010), but still don't obtain the ground truth communities. MaxBic and MaxBicR(0.6) outperform BiTector, which also merges maximal bicliques (one of its 4 communities is our community 5 and, similarly, is their only one split across ground truth communities) and linkcomm, the only algorithm based on edge similarity rather than node similarity. MaxBic is designed to find communities in the whole node set ($P \vee S$ not P), and is a first stage algorithm only, so finds more communities than the other algorithms.

Second, we consider published results where all nodes within the Southern Women network are clustered. Because there is no ground truth published for overlapping communities in the whole network and we have demonstrated that our communities well represent the ground truth, we make the assumption that the strongest structural communities (our 16 maximal bicliques in Table 1) are a true base level of the hierarchical overlapping community structure of this network. The community number and *NMI* for each algorithm is shown in Table 3.

Now we see that MaxBicR(0.6) has the highest *NMI* against MaxBic, as could be hoped, since we are merging the maximal bicliques found by MaxBic. It is encouraging that BiTector, which also merges maximal bicliques, detects the next most similar communities to MaxBicR. linkcomm performs next best, then COPRA (when preset to 2 overlapping communities). It may be reasonable to conclude that the Probabilistic model (preset to 2 overlapping communities) and BiLPA (which obtains overlapping communities if $\theta \leq 0.8$) are missing some fundamental community structure carried by maximal bicliques.

Table 3 Comparisons of MaxBic metadata communities with MaxBicR(0.6) and five other methods for detecting overlapping communities of Women and Events ($P \vee S$)

Method	No. of communities	*NMI*
MaxBic	16	1
MaxBicR(0.6)	9	0.818521
BiTector (Du et al. 2008)	4	0.679316
linkcomm (Ahn et al. 2010)	7	0.468365
COPRA (Gregory 2010)	2	0.452535
Probabilistic model (Chang and Tang 2014)	2	0.316433
BiLPA (Li et al. 2016)	4	0.315424

We will only continue comparisons of MaxBic and MaxBicR(0.6) with the two concurrent overlapping community detection algorithms COPRA and linkcomm, as they use methodologies different from MaxBic and BiTector. We tested COPRA for $v = 1, 2, \ldots, 10$ and report only the highest *NMI* obtained (which always occurred for $v = 2$ or 3).

Noordin top terrorist network

The Noordin Top terrorist network described in detail by the International Crisis Group (2006) was subsequently formalised in terms of several unipartite and bipartite networks. One of these is an affiliation bipartite network described in Roberts and Everton (2011); Everton (2012); Alzahrani and Horadam (2016) which links 74 individuals (set P) who had 45 affiliations (set S). Here $n = 119$ and $m = 276$. In Fig. 3 we plot the node degrees of the network.

Metadata analysis (Everton (2012), using information in International Crisis Group (2006)) partitioned set S into six categories, which we used in the following numerical order: Organizations (8 events), Operations (5), Training (11), Meeting (12), Finance (2) and Logistics (7).

In our previous study (Alzahrani and Horadam 2014) which partitioned the projected actor network G_P into 5 disjoint communities using *Infomap*, we noted that each actor could belong to more than one community. Applying MaxBic to the bipartite network results in 39 base level communities, of which 8 are almost strong, 23 are weak and 8 are very weak. Again, there are no strong communities, though all communities are maximal bicliques (possibly with some degree 1 nodes included). We list the almost strong communities in Table 4 and very weak communities in Table 5, as their small numbers are easier to visualise and evaluate. The weak communities are listed in Table 9 in the Appendix A.

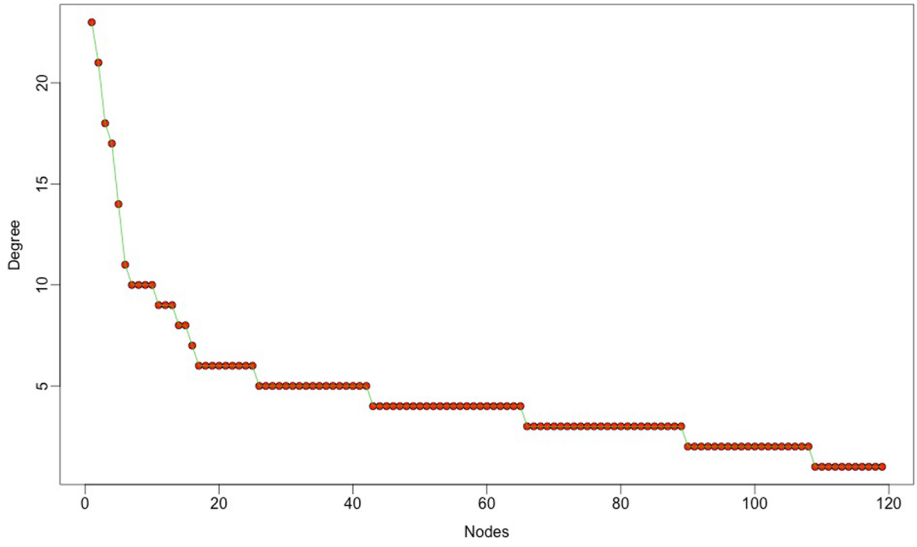

Fig. 3 Node degrees of the Noordin Top terrorist network. Noordin Top has degree 23, Organization 5 (Affiliation 5–Jemaah Islameyah) has 21, Operation 1 (Affiliation 9–Australian Embassy bombing Sep 2004) has 18 and Azhari Husin has 17. The two top actors under different centrality measures of the network (see (Roberts and Everton 2011; Alzahrani and Horadam 2014)), Noordin Top and Azhari Husin, stand out

Table 4 The 8 almost strong communities identified using MaxBic in Noordin Top terrorist network

No.	Almost strong communities	St
1	7 22 Abdullah Sunata Aris Munandar Asep Jaja Dani Chandra Hari Kuncoro	21
2	8 9 17 21 Apuy Fathurrochman Heri Golun Iqbal Huseini Iwan Dharmawan	16
3	8 9 15 Fathurrochman Heri Golun Iqbal Huseini Irun Hidayat Iwan Dharmawan	10
4	6 7 22 Abdullah Sunata Aris Munandar Asep Jaja Iqbal Huseini Umar Wayan	10
5	30 Joni Achmad Fauzan Musab Sahidi Said Sungkar Usman bin Sef	8
6	20 Enceng Kurnia Harun Hence Malewa	6
7	2 Abu Bakar Baasyir Adung Zulkarnaen	6
8	38 Ubeid	2

The smallest community is almost strong community 8 (see Table 4) of two nodes, *Ubeid* and Affiliation 38 (Finance 2). This is an example of a community that is not merged in Phase (vi) of MaxBic because its $|\text{list}| < 2$. For deeper analysis, we return to the metadata (International Crisis Group 2006) and the 6 affiliation categories in Everton (2012) to illustrate that our structural communities are meaningful in real terms.

The almost strong communities are visualised in Fig. 4. What is notable is that they do not contain any of the 5 most central individuals previously identified (Roberts and Everton (2011), Table 3). This is consistent with a decentralised cell structure in which the "footsoldiers", who are linked by Training events or belonging to the same Organisation, are not in direct contact with the network leaders.

The very weak communities are visualised in Fig. 5. By contrast, all but one of them contain one or both of the two most central individuals, *Noordin Top* (who ran the group) and his master bomb maker *Azhari Husin*, who frequently travelled together. In the main they are linked through internal communication (Meeting events) and Logistics.

In the weak communities (see Table 9 in Appendix A), for instance, community 16 has strong relationships between its members, because they were involved in virtually every major bombing (Operations). *Ali Ghufron* is the bomber in 2002 Bali Bombing I and collaborated with *Hambali*. In community 11, *Abdul Rauf, Imam Samudra* and *Iqbal* were also involved directly in the Operation of Bali Bombing I. The basis of community 3 is that its members have Trained together and were involved directly in the 2004 Australian Embassy Bombing Operation.

Overlap between our communities also has meaning. For instance, *Abdul Malik* and *Umar Wayan* overlap in weak communities 4 and 1 because they attended the same Organization (Jemaah Islamiyah) with other people in both communities such as *Abdullah*

Table 5 The 8 very weak communities identified using MaxBic in Noordin Top terrorist network

No.	Very weak communities
1	5 22 Fathurrahman al-Ghozi Hari Kuncoro Muchtar Umar Wayan
2	5 11 18 42 Azhari Husin Fathurrahman al-Ghozi Jabir
3	5 11 31 Azhari Husin Fathurrahman al-Ghozi Subur Sugiarto
4	5 18 Azhari Husin Cholily Fathurrahman al-Ghozi Jabir
5	34 35 41 Abu Dujanah Azhari Husin Noordin Top Qotadah
6	13 26 41 Azhari Husin Ismail Mohamed Rais Noordin Top
7	31 41 42 Ahmad Rofiq Ridho Azhari Husin Imam Bukhori Noordin Top
8	31 41 45 Azhari Husin Cholily Noordin Top

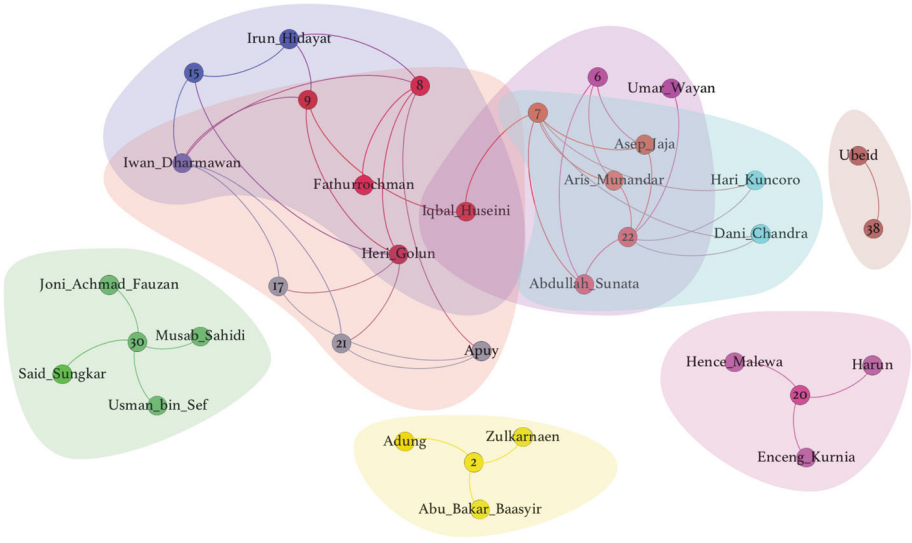

Fig. 4 Almost strong communities found by MaxBic

Sungkar and *Agus Ahmad*. *Apuy* overlaps in almost strong community 2 and weak community 3, reflecting the common Training category he attended with the members from both communities.

The actors with the most community memberships are the three who together form weak community 2: *Iqbal Huseini, Azhari Husin* and *Noordin Top* with 12, 11 and 10 memberships respectively. This indicates that they are important nodes in the network. In term of core and peripheral nodes, using the formulas (11) and (12) we have $\mu \approx 2.38$ and $\sigma \approx 2.32$, and on selecting $t_p = 1$ (to isolate nodes with membership 1), and $t_c = 1$ for consistency, we have $\mu + \sigma = 4.7$ and $\mu - \sigma = 0.06$. Therefore, the core nodes are those with membership 5 or more and peripheral nodes are those with membership 1.

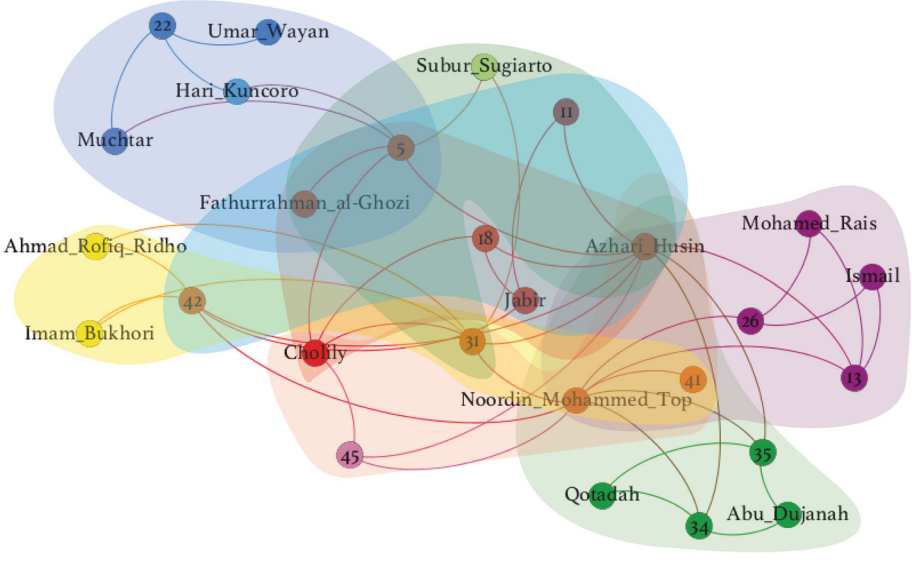

Fig. 5 Very weak communities found by MaxBic

Comparison with the other algorithms, taking MaxBic communities as metadata communities, appears in Table 6. We see that relative performance mimics that in Table 3, with MaxBicR(0.6) and linkcomm finding similar numbers of communities, more than COPRA, with NMI descending in the same order.

NSW crime network

This historic crime data from the Australian state of New South Wales (NSW) was published in 2013 (NSW Bureau of Crime Statistics and Research). It was collected from January 1995 to 2009, and includes data about every crime by month of occurrence, categorised by offence type. There are 21 offence categories (set S), some of which have subcategories, e.g. the category Homicide has four subcategories (Murder, Attempted Murder, Accessory to Murder and Manslaughter). The underlying social network of offenders is reflected in the reported crimes. The data reports the crime according to the local government area (LGA) in which it was committed (set P, $k = 155$). Here $n = 176$, $m = 8,761$, so we have a denser network than in the previous examples. No detection of overlapping communities has been undertaken to date.

In Alzahrani and Horadam (2014) using the partitioning algorithm *Louvain* on the projected network for P, no communities are detected, whereas using *Infomap* we found P partitioned into 2 communities, one (IC1) containing 82 LGAs and the other (IC2) containing 73 LGAs. When the LGAs are coloured on a map of NSW according to their *Infomap* community membership, a very strong geographical divide is apparent (see Fig. 6). Generally speaking, IC1 includes the more populated LGAs and IC2 includes the majority of rural and "Outback" LGAs. The 38 LGAs in the main metropolitan area, Sydney, are all in IC1. This provides external validation of the *Infomap* partition, and leads us to expect at least 2 overlapping communities in $P \vee S$.

Applying MaxBic to the NSW crime network results in 50 overlapping communities, of which 30 are strong and 20 are almost weak, according to Definition 1.

Comparison of MaxBicR with the two concurrent overlapping community detection algorithms, again assuming the MaxBic communities are metadata communities, appears in Table 7. MaxBicR(0.6) finds 8 communities, all of which are strong according to Definition 1. Neither linkcomm nor COPRA can detect any communities, contrary to the geographical evidence, and the evidence from MaxBic that there are 30 maximal bicliques that are strong communities within the network.

Closer analysis of the memberships of the 50 MaxBic communities, helps to reconcile these findings and understand this network better. In fact, 63 LGAs and 7 Offence categories have the maximum membership of 50, that is, they lie in *every* MaxBic community. This means these nodes form a $K_{63,7}$ biclique, which is the intersection of all 50 MaxBic

Table 6 Comparison of community numbers and NMI for Noordin Top terrorist network

Method	No. of communities	*NMI*
MaxBic	39	1
MaxBicR(0.6)	33	0.981
linkcomm (Ahn et al. 2010)	25	0.561785
COPRA (Gregory 2010)	6	0.435674

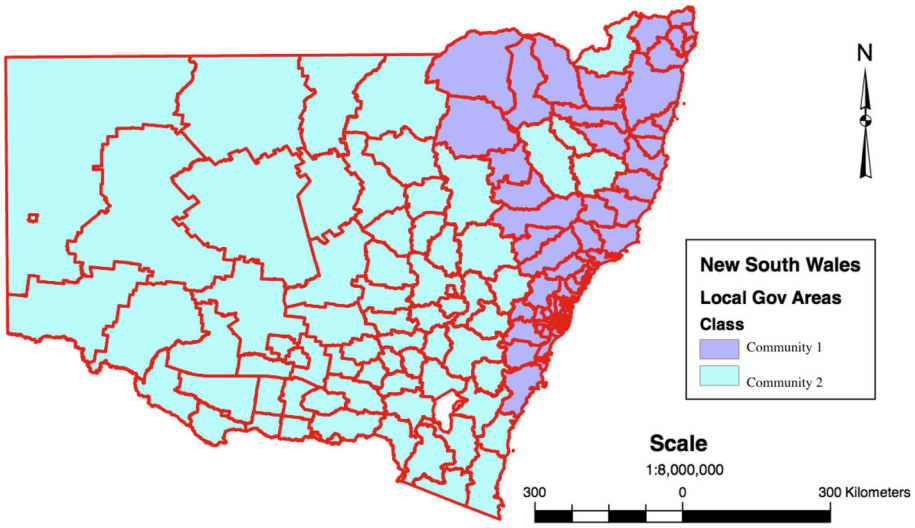

Fig. 6 Geographical divide of NSW communities found using *Infomap* (c.f. Alzahrani and Horadam (2014), Fig.1a). The uncoloured area is the Australian Capital Territory, which does not form part of NSW

communities and is a very dominant structure in the network. It is not, however, maximal, i.e. one of the 50 communities: every MaxBic community contains additional nodes from P or S. For example, only one of the 50 communities contains just the 63 common P nodes, but it contains all 21 S nodes.

To determine core and peripheral nodes, using the formulas (11) and (12) we have $\mu \approx 35.1$ and $\sigma \approx 14.23$, and on selecting $t_p = t_c = 1$ for consistency with the other datasets, we have $\mu+\sigma = 49.31$ and $\mu-\sigma = 20.85$. Therefore, the core nodes are (unsurprisingly) the 70 with membership 50 and peripheral nodes are those with membership 20 or less. There are 34 peripheral LGAs, of which all but 2 (both high socio-economic status Sydney LGAs) lie in IC2, and 3 peripheral Offence categories (Drug offences, Blackmail and extortion, and Prostitution offences).

We can conclude that MaxBic allows us to identify an extremely dominant structurally cohesive biclique in the network. Finally, in this dominant biclique the 63 LGAs are split, albeit very disproportionately, across the geographical divide, with 53 (84%) being in IC1 and 10 in IC2. This may account for the inability of the other algorithms to determine any communities, overlapping or otherwise.

Conclusion and future work

Bipartite networks are a very important class of complex networks, but have not received the same attention in community detection investigations as unipartite networks have. In

Table 7 Comparison of community numbers and NMI for NSW crime network

Method	No. of communities	*NMI*
MaxBic	50	1
MaxBicR(0.6)	8	0.68345
linkcomm (Ahn et al. 2010)	1	0.36458
COPRA (Gregory 2010)	1	0.36458

particular, there is no accepted standard planted community model for generating synthetic bipartite networks that can be used to compare performance of overlapping community detection algorithms. Consequently, at present, performance can fundamentally only be measured against ground truth or metadata analysis where this exists.

We have introduced a new maximal biclique-finding algorithm, MaxBic, for detection of overlapping communities in bipartite networks. It is based on node similarity and finds an order-limited number of overlapping maximal bicliques which (after allowing for some nodes of degree 1) cover an unweighted undirected connected bipartite network. We can categorise and rank these bicliques by new measures of relative strength.

We have shown MaxBic has time complexity at worst $O\left(n^2 \min\{k, l\}\right)$, irrespective of whether the network is sparse or dense, whereas the problem of finding all maximal bicliques runs at least exponentially in n. The improvement occurs because MaxBic finds at most n maximal bicliques.

The overlapping community structure produced by our algorithm consists of maximal bicliques, with perhaps some degree 1 nodes included. At a formal level therefore, it represents communities that are as strong as possible in a bipartite network representation and so truly captures structural communities that can form the base level of a community hierarchy. We tested this base level for three small bipartite networks for which some ground truth or metadata analysis was available, in some detail. We conclude that the overlapping communities we find do capture ground truth in a real sense.

Whilst the base level of the overlapping community hierarchy carries important information, in large dense networks it may be too much information for sensible analysis. Most effective algorithms use further techniques to reduce the number of communities found by amalgamating smaller ones according to some measure of community strength. We introduce a second stage algorithm which will merge communities if their fraction of overlap passes a fixed threshold (here set at 0.6) and use it to show that the communities it finds, at the next level of the hierarchy, better capture the maximal biclique information present in the data than do competing algorithms. Any other second stage algorithm could be applied and further work could compare performance of several of these when input with the MaxBic output communities.

Analysis of overlapping communities in the Noordin Top terrorist network and other dark networks should bring new insights. We can identify core actors and most common affiliations. For instance, in this kind of terrorist network, hidden relations can be observed through overlapping communities, and actors who have more connections and overlap with many others might have more influence and may be more dangerous persons.

Similarly, deeper analysis of the strong communities in the NSW crime network may help identify or confirm gang locations and offences they specialise in. To counteract the dominance of the intersection biclique we discovered, it may be useful to include all the offence subcategories in further experiments.

Finally, we aim to speed up the running time of MaxBic by further optimizing our code. A slow point is Phase (iv-v) but this could perhaps be implemented to run more quickly on real datasets by e.g. comparing only those node pairs which are more likely to be similar, not all of them.

Appendix A

Algorithm 1 MaxBic: Phases (i-vi)

Input: E (edge list)
$G = (P, S, E)$
$n = |P \vee S|$ //order of G
$k = |P|$
for $i = 1$ to k **do**
    $\Gamma(p_i) = \{y \in S : edge(p_i, y) \in E\}$
end for
for $i = 1$ to k **do**
    for $j = i + 1, j > i$ to k **do**
        $CNS_{ij} = \Gamma(p_i) \cap \Gamma(p_j)$
        if $CNS_{ij} \neq \emptyset$
        Clique $K^*(p_i, p_j) \Longleftarrow CNS_{ij} \cup p_i \cup p_j$ // assign nodes to new basic clique
        for $i^* = 1$ to $|K^*(p_i, p_j)|$ **do**
            for $j^* = i^* + 1, j^* > i^*$ to $|K^*(p_i, p_j)|$ **do**
                $E^* \Longleftarrow edge(i^*, j^*)$ // insert edges to E^*
                $G^* = (P, S, E^*)$
            end for
        end for
    end for
end for
for $i = 1$ to n **do**
    $c_i = \{i :\} \Longleftarrow i$ // initial step in $G*$, take each node and assign to unique cluster
    $\Gamma^*_+(i)$ // find inclusive neighbors for each cluster
end for
for $i = 1$ to n **do**
    $c_i = main$ // main identifies the cluster we compare with other clusters
    for $j = 1, j \neq i$ to n **do**
        if $\Gamma^*_+(i) \subseteq \Gamma^*_+(j)$
        $c_i \Leftarrow c_j$ // join cluster c_j to cluster c_i
        $Z \Leftarrow \{i{:}j \ldots\}$ // Z is a container which contains node i and its joined nodes
    end for
end for
Return: $Z = \{c_i\}$ // vector of sets of form $c_i = \{i : j_1, j_2 \ldots\}$; i is the element and $j_1, j_2 \ldots$ is the list
$F = |Z|$
for $i = 1$ To F **do**
    for $j = 1, j \neq i$ to F **do**
      if $c_i \subseteq c_j$
      remove c_i from Z
    end for
end for
// now we compare list of each set as total subset of another set in order to reduce set redundancy.
for $i = 1$ to F **do**
    for $j = 1, j \neq i$ to F **do**
      if $\text{list}_i \subseteq \text{list}_j$ AND $|list_i| \geq 2$ **Then**
        element $c_i \cup c_j$
    end for
      if $|list_i| = 1$
    $c_i \cup \Gamma(c_i)$
end for

Algorithm 1 MaxBic Phase (vii): Order communities by strength

$F = |Z|$
for $i = 1$ To F **do**
  $c_i \Longleftarrow i$
  for $j = 1$ To $|c_i|$ **do**
    $k_j^{in} = \sum_{x \in c_i} a_{xj}$
    $k_j^{out} = \sum_{x \notin c_i} a_{jx}$
  end for
end for
for $i = 1$ To F **do**
  $c_i \Longleftarrow i$
  for $j = 1$ To $|c_i|$ **do**
    for $k = 1$ To F **do**
      $k_j^{max-out} = max_{c_k \neq c_i} \sum_{x \in c_k} a_{jx}$
    end for
  end for
end for
for $i = 1$ To F **do**
  $c_i \Longleftarrow i$
  for $j = 1$ To $|c_i|$ **do**
    if $k_j^{in} > k_j^{out} \quad \forall j \in c_i,$ **then**
    $Strong\_community \Longleftarrow c_i$ // assign node to strong community
    Else if $k_j^{in} \geq k_j^{max-out} \quad \forall j \in c_i,$ **then**
    $Almost\_strong\_community \Longleftarrow c_i$ // assign node to almost strong community
    Else if $\sum_{j \in c_i} k_j^{in} > \sum_{j \in c_i} k_j^{out}$ **then**
    $Almost\_weak\_community \Longleftarrow c_i$ // assign node to almost weak community
    Else if $\sum_{j \in c_i} k_j^{in} \geq \sum_{j \in c_i} k_j^{max-out}$ **Then**
    $Weak\_community \Longleftarrow c_i$ // assign node to weak community
    Else
    $Very\_weak\_community \Longleftarrow c_i$ // assign node to very weak community
  end for
end for
for $i = 1$ To F **do**
  $c_i \Longleftarrow i$
  if $c_i \subseteq Strong\_community$
  $St(c_i) = \sum_{j \in c_i} k_j^{in} - \sum_{j \in c_i} k_j^{out}$
  Else if $c_i \subseteq Almost\_strong\_community$
  $St(c_i) = \sum_{j \in c_i} k_j^{in} - \sum_{j \in c_i} k_j^{max\text{-}out}$
  Else if $c_i \subseteq Almost\_weak\_community$
  $St(c_i) = \sum_{j \in c_i} k_j^{out} - \sum_{j \in c_i} k_j^{in}$
  Else if $c_i \subseteq Weak\_community$
  $St(c_i) = \sum_{j \in c_i} k_j^{max\text{-}out} - \sum_{j \in c_i} k_j^{in}$
  Else
  $Very\_weak\_community \Longleftarrow c_i$
end for
Sort(*Strong_communities*)
Sort(*Almost_strong_communities*)
Sort(*Almost_weak_community*)
Sort(*Weak_community*)
result $\Longleftarrow$ ordered strong and almost strong communities by **larger** value
result $\Longleftarrow$ ordered weak and almost weak communities by **smaller** value
Return: result community file

Further data and results

Table 8 Representation of the adjacency matrix of Southern Women bipartite network, from (Freeman 2003)

	Event	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Woman															
1	Evelyn	1	1	1	1	1	1		1	1					
2	Laura	1	1	1		1	1	1	1						
3	Theresa		1	1	1	1	1	1	1	1					
4	Brenda	1		1	1	1	1	1	1						
5	Charlotte			1	1	1		1							
6	Frances			1		1	1		1						
7	Eleanor					1	1	1	1						
8	Pearl						1		1	1					
9	Ruth					1		1	1	1					
10	Verne							1	1	1			1		
11	Myra								1	1	1		1		
12	Katherine								1	1	1		1	1	1
13	Sylvia							1	1	1	1		1	1	1
14	Nora						1	1		1	1	1	1	1	1
15	Helen							1	1		1	1	1		
16	Dorothy								1	1					
17	Olivia									1		1			
18	Flora									1		1			

Table 9 The 23 weak communities identified using MaxBic in Noordin Top terrorist network

No.	Weak communities	St
1	1 8 9 Abdul Malik Agus Ahmad Ajengan Masduki Akram Engkos Kosasih Fathurrochman Iqbal Huseini Irun Hidayat Umar Wayan	-13
2	9 11 13 26 31 32 34 35 36 41 42 45 Azhari Husin Iqbal Huseini Noordin Top	-12
3	9 17 Apuy Baharudin Soleh Heri Golun Iqbal Huseini Iwan Dharmawan Umar	-11
4	1 5 Abdul Malik Abdullah Sungkar Ajengan Masduki Akram Chandra Engkos Kosasih Fathurrahman al-Ghozi Umar Wayan	-10
5	8 9 15 17 21 Fathurrochman Heri Golun Iqbal Huseini Iwan Dharmawan	-10
6	5 13 Asmar Latin Sani Azhari Husin Fathurrahman al-Ghozi Ismail Mohamed Ihsan	-10
7	3 9 33 37 41 Abu Fida Iqbal Huseini Son Hadi	-9
8	29 31 32 41 42 43 44 Ahmad Rofiq Ridho Noordin Top	-9
9	9 19 27 Aceng Kurnia Achmad Hasan Heri Sigu Samboja Iqbal Huseini	-8
10	6 7 22 40 Abdullah Sunata Asep Jaja Iqbal Huseini Umar Wayan	-7
11	8 10 Abdul Rauf Fathurrochman Imam Samudra Iqbal	-6
12	21 39 Iwan Dharmawan Saptono Urwah	-6
13	8 16 Fathurrochman Irun Hidayat Iwan Dharmawan Rosihin Noor	-5
14	11 24 41 Anif Solchanudin Misno Noordin Top Salik Firdaus	-5
15	3 25 Achmad Hasan Son Hadi Suramto	-5
16	4 10 Ali Ghufron Hambali Marwan	-4
17	5 10 12 13 Azhari Husin Fathurrahman al-Ghozi Mohamed Ihsan Toni Togar	-3
18	32 41 44 Ahmad Rofiq Ridho Joko Triharmanto Noordin Top Purnama Putra	-3
19	5 10 14 Dulmatin Fathurrahman al-Ghozi Marwan Umar Patek	-3
20	13 26 28 41 Ismail Mohamed Rais Noordin Top	-2
21	7 32 41 44 Iqbal Huseini JokoTriharmanto Purnama Putra	-2
22	10 12 Azhari Husin Hambali Imam Samudra Mohamed Ihsan Toni Togar	-1
23	23 34 35 41 Noordin Top Qotadah	0

Acknowledgements
Not applicable.

Funding
Not applicable.

Availability of data and materials
The Southern Women dataset analysed in this study is published in Davis et al. (1941). The Noordin Top dataset (International Crisis Group 2006) analysed in this study is available from T.A. on reasonable request. The NSW crime dataset (NSW Bureau of Crime Statistics and Research) analysed in this study is available at http://data.gov.au/dataset/nsw-crime-data/.

Author's contributions
Part of this work formed part of the PhD research of T.A. taken under the supervision of K.H.. T.A. conceived and implemented the algorithms and performed the experiments; T.A. and K.H. designed the experiments, analyzed the data and wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 15 November 2018 Accepted: 12 March 2019
Published online: 02 May 2019

References
Ahn Y-Y, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466:761–764
Alexe G, et al (2004) Consensus algorithms for the generation of all maximal bicliques. Discret Appl Math 145:11–21
Alzahrani T (2016) Complex information networks – detecting community structure in bipartite networks. PhD Thesis, RMIT University. Australia
Alzahrani T, Horadam KJ (2014) Analysis of two crime-related networks derived from bipartite social networks. In: Proceedings of 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE. pp 890–897
Alzahrani T, Horadam KJ (2016) Community Detection in Bipartite Networks: Algorithms and Case studies. In: Lü J, Yu X, Chen G, Yu W (eds). Complex Systems and Networks: Dynamics, Controls and Applications. Springer Berlin Heidelberg. pp 25–50
Barber MJ (2007) Modularity and community detection in bipartite networks. Phys Rev E 76(6):066102
Barrat A, Barthelemy M, Vespignani A (2008) Dynamical processes on complex networks. Cambridge University Press, Cambridge
Blondel V, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 10:P10008
Cazabet R, Amblard F, Hanachi C (2010) Detection of overlapping communities in dynamical social networks. In: second international conference on social computing. IEEE. pp 309–314
Chang C, Tang C (2014) Community detection for networks with unipartite and bipartite structure. New J Phys 093001:16
Cui Y, Wang X (2014) Uncovering overlapping community structures by the key bi-community and intimate degree in bipartite networks. Physica A: Stat Mech Appl 407:7–14
Davis A, Gardner BB, Gardner MR (1941) Deep south: A Social Anthropological Study of Caste and Class. University of Chicago Press, Chicago
Du N, Wang B, Wu B, Wang Y (2008) Overlapping community detection in bipartite networks. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. IEEE. pp 176–79
Esquivel A, Rosvall M (2011) Compression of flow can reveal overlapping-module organization in networks. Phys Rev X 1(2):021025
Evans TS (2010) Clique graphs and overlapping communities. J Stat Mech Theory Exp 2010(12):P12037
Everton SF (2012) Disrupting Dark Networks. Cambridge University Press, Cambridge
Freeman LC (2003) Finding social groups: A meta-analysis of the southern women data. In: Dynamic social network modeling and analysis. National Academies Press. pp 39–97
Fortunato S. (2010) Community detection in graphs. Phys Rep 486:75–174
Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99:7821–7826
Gregory S (2010) Finding overlapping communities in networks by label propagation. New J Phys 103018:12
Hu Y, Chen H, Zhang P, Li M, Di Z, Fan Y (2008) Comparative definition of community and corresponding identifying algorithm. Phys Rev E 78(2):026121
International Crisis Group (2006) Terrorism in Indonesia: Noordin's Networks. Asia Report no. 114. International Crisis Group, Brussels
Kalinka AT, Tomancak P (2011) linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. Bioinformatics 27(14):2011–2012

Lancichinetti A, Fortunato S (2014) Erratum to: Community detection algorithms: a comparative analysis. [Physical Review E, 80, 5, 056117, 2009]. Phys Rev E 89(5):049902
Lancichinetti A, Fortunato S, Kertész J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New J Phys 11(3):033015
Lázár A, Ábel D, Vicsek T (2010) Modularity measure of networks with overlapping communities. EPL (Europhys Lett) 90(1):18001
Larremore D, Clauset A, Jacobs A (2014) Efficiently inferring community structure in bipartite networks. Phys Rev E 90(1):012805
Lehmann S, Schwartz M, Hansen LK (2008) Biclique communities. Phys Rev E 78(1):016108
Leicht E, Holme P, Newman ME (2006) Vertex similarity in networks. Phys Rev E 026120:73
Leung IXY, Hui P, Lio P, Crowcroft J (2009) Towards real-time community detection in large networks. Phys Rev E 79(6):066107
Li Z, Wang RS, Zhang S, Zhang XS (2016) Quantitative Function and Algorithm for Community Detection in Bipartite Networks. Inf Sci 367-368:874–889
Liben Nowell D, Kleinberg J (2007) The link prediction problem for social networks. J Am Soc Inf Sci Technol 58:1019–1031
Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. J Math Sociol 1:49–80
Luce RD, Perry AD (1949) A method of matrix analysis of group structure. Psychometrika 14(2):95–116
Makin K, Uno T (2004) New Algorithms for Enumerating All Maximal Cliques(Hagerup T, Katajainen J, eds.) SWAT 2004, LNCS 3111
Moody J, White DR (2003) Structural cohesion and embeddedness: A hierarchical concept of social groups. Am Sociol Rev 68(1):103–127
Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103:8577–8582
Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 026113:69
NSW Bureau of Crime Statistics and Research NSW Crime data. Historic. Published 2013. http://data.gov.au/dataset/nsw-crime-data/. Accessed Mar 2012
Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814–818
Peeters R (2003) The maximum edge biclique problem is NP-complete. Discret Appl Math 131(3):651–654
Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Natl Acad Sci 101:2658–2663
Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 036106:76
Roberts N, Everton SF (2011) Strategies for combating dark networks. J Soc Struct 12:2
Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105:1118–1123
Ruspini EH (1970) Numerical methods for fuzzy clustering. Inf Sci 2(3):319–350
Tarissan F (2015) Comparing overlapping properties of real bipartite networks. In: Interdisciplinary Symposium on Complex Systems. Springer. pp 309–17
Van Steen M (2010) Graph theory and complex networks. An Introduction. Lexington
Viard T, Latapy M, Magnien C (2016) Computing maximal cliques in link streams. Theor Comput Sci 609(1):245–252
Xie J, Kelley S, Szymanski BK (2013) Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Comput Surv (CSUR) 45(4):43
Xu Y, Chen L, Zou S (2013) Community detection from bipartite networks. In: 10th Web Information System and Application Conference. IEEE. pp 249–254
Zhou T, Lü L, Zhang YC (2009) Predicting missing links via local information. Eur Phys J B 71(4):623–630