Community Detection in Bipartite Networks Using Random Walks

Taher Alzahrani, Kathy J. Horadam, and Serdar Boztas

Abstract. Community detection plays a crucial role in many complex networks, including the increasingly important class of bipartite networks. Modularity-based community detection algorithms for bipartite networks are hampered by their well-known resolution limit. Unfortunately, the high-performing random walk based algorithm Infomap, which does not have the same constraint, cannot be applied to bipartite networks.To overcome this we integrate the projection method for bipartite networks based on common neighbors similarity into Infomap, to acquire a weighted one mode network that can be clustered by the random walks technique. We also compare results obtained from this process with results in the literature. We illustrate the proposed method on four real bipartite networks, showing that the random walks technique is more effective than the modularity technique in finding communities from bipartite networks as well.

Keywords: bipartite graph, community detection, random walks.

1 Introduction

One mode, or unipartite, networks are the typical framework for complex networks. Many techniques have been constructed to analyze them. However, many complex networks can best be described as bipartite [1]. A bipartite network is a network in which there are two different types of nodes, and the edges between nodes may occur only if nodes belong to different types. In the last few years, there has been increasing motivation to analyse bipartite networks as a separate network category, and in particular to investigate their community structure. For unipartite networks,

Taher Alzahrani · Kathy J. Horadam · Serdar Boztas
School of Mathematical and Geospatial Sciences,
RMIT University
Melbourne 3001, Australia
e-mail: {taher.alzahrani,kathy.horadam,serdar.boztas}@rmit.edu.au

P. Contucci et al. (eds.), *Complex Networks V*,
Studies in Computational Intelligence 549,
DOI: 10.1007/978-3-319-05401-8_15,

two approaches to community detection have been very popular, one based on modelling the community structure and one based on extracting it from flow calculations on the network. The best algorithms to cluster very large networks using each approach [2], compared using the LFR benchmark datasets [3], are now referred to as the Louvain algorithm [4] and the Infomap algorithm [5]. Unfortunately it is impossible to reformulate Infomap on a bipartite network, since then there is no stationary distribution for the probability of the walker to be at a given node on the bipartite graph. In other words, if you start in one class (set), then you will always be in that class after an even number of steps, so the probability of being at a particular vertex is zero at odd time steps. In the language of Markov chains, the random walk on a bipartite graph is periodic. The focus of this paper is on community detection in the weighted one mode networks which are projected from unweighted bipartite networks. This is a natural focus, as usually one set of nodes in a bipartite network, denoted the *primary set* P is of more interest for a particular purpose than the other, the *secondary set* S. The rôles of the two node sets can be switched for different applications. Our contribution is to apply a random walks based algorithm to the unipartite network projected on the primary set of the bipartite network. We also compare our results with the results in the literature that used bipartite modularity based algorithms. We investigate the communities found by Infomap in one case in detail, to demonstrate that the small communities found (below the resolution limit of modularity-based algorithms) represent real information.

2 Previous Work

Identifying communities, also called modules or clusters, in a network allows us to explore its hierarchical structure. This leads to better understanding of the major functions of the network, and more efficient spread of ideas, goods or services in the network.

2.1 Unipartite Networks

Girvan and Newman [6] initiated recent work on defining and evaluating communities, introducing the fast greedy technique which relies on a quality function called modularity. Modularity is a scalar measure of the quality of modules extracted from a network. The partition which maximises it is regarded as the best. The complexity of the algorithm is $O(n^3)$, where n is the number of nodes. Since the limited nodes size for previous algorithm is 10^3 many efforts have been devoted to upgrade the computational time of modularity optimization. For instance, the Radicchi et al [7] algorithm is in spirit the Girvan-Newman method but it iteratively removes the edges with highest clustering coefficient instead of edges with highest betweenness. The stated complexity of this algorithm is $O(n^2)$. Another algorithm that takes modularity optimization as its main quality function is that of Guimera and Amaral [8]. The fast modularity optimization algorithm by Blondel et al [4] has the best results compared with the previous algorithms. It is described as a multi level method

in community detection. The complexity of the Louvain algorithm is linear in the number of edges in the network, that is $O(m)$. However, modularity-based algorithms have a known drawback: a resolution limit in detecting communities [9]; that is, communities with internal edge numbers $\leq O(\sqrt{m})$ cannot always be reliably detected. All these methods, which rely mainly on modularity, describe and reveal communities in networks according to how the networks are built or by modelling their structure. However, a different method using random walks, known as Infomap proposed by Rosvall and Bergstrom [5], identifies communities according to information flow in the networks. The quality function used, called the map equation, is based on minimum description length (MDL) [10]. This function measures the average length $L(M)$ in bits per step of a random walk on a network with the modular partition M, as follows:

$$L(M) = q_{\curvearrowright} H(Q) + \sum_i p^i_{\circlearrowright} H(P^i) \tag{1}$$

where $q_{\curvearrowright}$ is the probability that the random walk moves between modules, $H(Q)$ is the entropy of module names, $p^i_{\circlearrowright}$ the probability of movement within module i and $H(P^i)$ is the entropy of the movements within module i. The complexity of the Infomap algorithm is also $O(m)$. However, Infomap does not apply to bipartite networks because a stationary distribution cannot be determined in general.

2.2 Bipartite Networks

Most authors follow Newman and Girvan's modularity method [6] to determine communities in bipartite networks. Michel et al [11] used unipartite Girvan-Newman modularity [12] as their standard model to derive a bipartite modularity model by building an unweighted "biadjacancy matrix" of a bipartite graph. Guimera [13] produces a modularity measurement for bipartite networks, denoting the sets of this network as actors and teams. The emphasis on finding modules here was in the actor set (P) after projection, with projection based on joint participation in teams. In [14], Barber developed a modularity matrix for bipartite networks, inspired by Newman's modularity matrix [15]. The previous approaches to finding modularity in bipartite networks were extended from Newman modularity [6]. A fast technique for unipartite networks, the Label Propagation Algorithm (LPA) [16], uses the local network structure as a guide for finding communities very efficiently (almost linearly in m). Barber and Clark [17] introduced a version of LPA, denoted LPAb, for bipartite networks. The speed of LPAb (complexity near linear in total number of edges m) makes it comparable with the fastest bipartite modularity optimization algorithm [18]. However, Liu and Murata [18] introduce a new version of LPAb, denoted LPAb+, which they claim as the most reliable algorithm with the highest bipartite modularity.

3 Method

The Infomap algorithm utilizes the information flow on a network in order to achieve its clustering. This information flow is approximated in practice by means of a random walk along the network, and iterating until a steady state distribution emerges, as it must, under the assumption that the network in question is strongly connected and aperiodic. Unfortunately for us, in general there is no stationary distribution of a random walk on bipartite networks that can be found from power iterations as discussed in Section 1. Thus Infomap cannot be directly used for our problem. Instead, we apply a projection method based on common neighbors similarity for our bipartite networks. The motivation is to be able to obtain a stationary distribution for the walk on the nodes in the unipartite network obtained by projection. This is achieved by integrating the projection process into the Infomap and Louvain algorithms. This allows us to compute the complexity time for the whole operation starting from converting bipartite networks to weighted unipartite networks followed by clustering them by the two algorithms. The reason the projection method is also applied to the Louvain algorithm is to be able to compare the performance of Infomap with that of Louvain, for the bipartite network case. The lack of existing benchmark bipartite networks motivated our work. Moreover, there is no evaluated community detection method in the literature that examines bipartite networks from a random walks perspective. We have programmed our projection algorithm in C++ for compatibility with the implementations we have of the Infomap and Louvain algorithms. We start by reading the bipartite network as a pair of nodes, the first from P and the second from S. The labels on the nodes in this dataset do not have to be numbers, they can be post codes, book serials, bank card numbers, names of social networks or even names of people. Then, we find the common neighbors between nodes i and j in P according to the following adjacency matrix A_{ij}:

- $A_{ij} = 1$, if nodes i and j have a common neighbor
- $A_{ij} = 1$, if node i has a neighbor which has no other neighbors in P (resulting in self loop, i=j)
- $A_{ij} = 0$, Otherwise

The weight of the edge between i and j in the projected unipartite network is the number of common neighbors of node i and node j.We also use special techniques in C++ that improved the efficiency of the projection method. Starting by using a C++ container called Mapvector which requests a key and a value, we choose the key to be the common neighbors and the value to be a vector of nodes $\{v_1, v_2, .., v_n\}$ where n is total number of nodes. Then, we create pairs in a one mode network and store the result in container called "Multiset".

4 Results and Discussion

Both algorithms were tested in four real world bipartite networks: the Southern women network, Newman's scientific collaboration network, a historical Australian

Table 1 Network sizes, where P and S are the number of primary set nodes and secondary set nodes respectively and m is the total number of edges

Network	P	S	m
Southern women	14	18	89
Scientific collaboration	16726	22016	58595
Australian government contracts	11924	1655	70019
NSW Crimes	155	22	9611

government contracts network and an Australian crime network. Their primary and secondary set sizes and total number of edges are given in Table 1.

Our results are summarised in Table 3. Since the Southern women network has been studied widely in the literature we investigate it in some detail first.

4.1 Southern Women Network

The "Southern women" network collected by Divas et al [19] has become a benchmark for testing community detection algorithms on bipartite networks. This network has 18 women (who form the primary set P) who attended 14 different events (the secondary set S). An edge exists between two women for each event they attend together. Most studies conducted before 2003 identify two (sometimes overlapping) communities of women while one identifies three communities [20]. In many studies, members within each community are further partitioned into core or peripheral members. More recent studies using bimodularity find more communities (3 and 4). Consequently, at least two communities are expected. Our implementation of the projection in Infomap produces 4 communities as shown in Figure 1. In Table 2, we list the community numbers found in the Southern women dataset by the more recent bipartite network algorithms described in Sections 2.2 and 3. We compare our results for the Southern women network with results in the literature, in more detail. Using Infomap, we have community A consisting of Evelyn and Theresa (women 1 and 3, respectively), community B consisting of Katherine and Nora (women 12 and 14, respectively), and two others $C = \{8, 9, 16, 17, 18\}$ and $D = \{2, 4, 5, 6, 7, 10, 11, 13, 15\}$, as shown in Figure 1. Our groups A and B consist of women frequently identified as core members of each of the two communities found in earlier studies. By contrast, Barber's two smaller communities consist of women who tended to be identified as peripheral members of each of the two communities found in earlier studies [20]. Barber also tested the success of his partition into four communities, found using the maximum bipartite modularity (as described in Section 2.2), as a partition in the corresponding *unweighted* projection network, and found it to have negative modularity [14]. As this is worse than considering the women as a single community, it further supports our use of the *weighted* projection network. Guimera et al [13] found only two communities of women (red and blue) whether modularity on the unweighted projection, the weighted projection or bipartite modularity was used. They found the

communities were inaccurate with unweighted projection, but identical and in agreement with supervised results in [20] for the other two methods. The total number of edges in the Southern women network after weighted projection is 139 edges. Our community A (Evelyn and Theresa) has 7 internal edges and lies inside the red group, while our community B (Katherine and Nora) has 5 internal edges and lies inside the blue group. These two "core" communities are not detected by the modularity based algorithm, probably because their edge numbers fall below the resolution limit of modularity [9] which in this case is 12 (since $11 < \sqrt{139} < 12$). By comparison the 2 communities found by our Louvain algorithm have 45 and 33 internal edges. This demonstrates that the resolution limit for modularity applies to Louvain but is passed by Infomap, in this benchmark bipartite case.

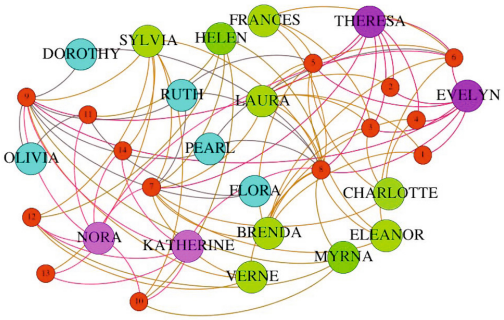

Fig. 1 The four communities of women found in the Southern women dataset. Red nodes represent S, the events the women attended, and the four other colors represent four communities within P, with nodes labelled by first name.

Table 2 Numbers of communities of women detected by different algorithms in the Southern women network

Algorithm	Quality function	Network applied to	Modules in P
Guimera [13]	modularity	weighted projection	2
Michel [11]	bimodularity	bipartite	3
Barber [14]	bimodularity	bipartite	4
LPAb(+) [18]	bimodularity	bipartite	4
This paper	map equation	weighted projection	4

4.2 The Scientific Collaboration Network

The scientific collaboration network in Newman [21], contains a bipartite network. It lists the relationships between publications and the scientists who are authors of these papers. There are 16726 scientists who wrote 22016 papers in this network. The number of edges between scientists and papers is 58595. The primary set in our projection is the scientists while the secondary set is the papers. Therefore, we are

Table 3 Community numbers obtained from our experiments, where L is the code length and Q is the modularity

	Infomap		Louvain	
Network	Comm.	L	Comm.	Q
Australian government contracts network	1114	8.340	836	0.530
Scientific collaboration network	2131	6.164	1266	0.877
Southern women network	4	3.992	2	0.352
Crime network	2	7.276	1	0.0

interested in detecting the communities of authors and determining who are more likely to collaborate together. The value of characterizing scientists in communities and describing the ties between distinct communities from different disciplines is important knowledge because it can help scientists collaborate. Our method utilizes a modified Infomap algorithm to characterize the scientific collaboration network into 2131 communities. The same method using Louvain algorithm finds fewer communities: 1266. We attribute these different results to the resolution limit of modularity optimization in Louvain algorithm. Scientists in one community have more in common than in another and they are likely to collaborate together and this increases the strength of the community.

4.3 The Historical Australian Government Contracts Network

The historical Australian government contract data has been published in 2012 [22]; it contains great detail about agencies and companies that undertake projects in Australia. We construct a bipartite network from this dataset. The network in this case has the ABN (Australian Business Number) a unique identifier number for agencies and companies, as the primary set. The postcode areas which these agencies have projects in them form the second set. The bipartite network has 11924 nodes as agencies and/or companies, and 1655 different postcode areas. The number of edges in this dataset is 70019, which is number of projects from 1999 until 2012 [22]. The weighted one mode projected network relates agencies that have common projects in the specific postcode area. The results produced from our method implemented in Infomap illustrate 1114 communities are found which contained agencies working on projects in the same postcode area. However, there were different results from Louvain where only 836 communities have been identified. The investigation of the historical Australian government contract data could lead to more collaboration between agencies/companies if they have projects in the same postcode area.

4.4 Crime Network

This monthly data is collected from January 1995 to December 2009, and shows the crimes and offences committed in New South Wales (NSW) in Australia [23].

Moreover, it provides the location of the crimes, thus we are interested in the location and the crime itself, which form our bipartite network. The primary set in this data is the location of the crime, whereas the offence is the secondary set. The intention in finding communities in the crime dataset is to identify and illustrate where similar crimes have occurred. There are 155 locations and 22 types of offences committed. The number of crimes committed during almost 14 years is 9611, which is the number of edges in this network. Results from applying our integrated Infomap algorithm show that there are two communities (one with 73 locations and the other with 82) where similar crimes are being committed. However, only one community that is the entire state of NSW is found when applying the Louvain algorithm, so this algorithm provides no useful information.

5 Conclusion

In this paper, we have integrated the projection method for bipartite networks into Infomap, to acquire a weighted one mode network that can be clustered by the random walks technique. The results from this process reflect valuable information compared with bipartite network based algorithms in the literature. Experiments on four real world bipartite networks show that a random walk based algorithm is more functional in detecting the communities in the primary set of a bipartite network than a modularity based algorithm. For future work, we intend to modify our approach to project the two sets P and S of a bipartite network in parallel, cluster them under random walks algorithms and then merge the whole into a clustered bipartite network. Moreover, weighted bipartite networks, overlapping communities, measuring quality of the communities found and difference of the quality between (projection + Infomap) and a method which would compute bipartite communities and then project the will be taken into consideration.

Acknowledgement. The first author would like to thank the Ministry of Finance of Saudi Arabia for supporting his research. The work of the second and third authors was partly supported by Department of Defence of Australia Agreement 4500743680.

References

1. Nishikawa, T., Motter, A.E., Lai, Y.C., Hoppensteadt, F.C.: Heterogeneity in oscillator networks: Are smaller worlds easier to synchronize? Physical Review Letters 91, 014101 (2003)
2. Lancichinetti, A., Fortunato, S.: Community detection algorithms: A comparative analysis. Physical Review E 80, 056117 (2009)
3. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Physical Review E 78, 046110 (2008)
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008 (2008)

5. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105, 1118–1123 (2008)
6. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69, 026113 (2004)
7. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of the United States of America 101, 2658–2663 (2004)
8. Guimera, R., Sales-Pardo, M., Amaral, L.S.A.N.: Modularity from fluctuations in random graphs and complex networks. Physical Review E 70, 025101 (2004)
9. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences 104, 36–41 (2007)
10. Grunwald, P.D., Myung, I.J., Pitt, M.A.: Advances in minimum description length: Theory and applications. MIT press (2005)
11. Crampes, M., Plantie, M.: A Unified Community Detection, Visualization and Analysis method. arXiv preprint arXiv:1301.7006 (2013)
12. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99, 7821–7826 (2002)
13. Guimera, R., Sales-Pardo, M., Amaral, L.S.A.N.: Module identification in bipartite and directed networks. Physical Review E 76, 036102 (2007)
14. Barber, M.J.: Modularity and community detection in bipartite networks. Physical Review E 76, 066102 (2007)
15. Newman, M.E.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74, 036104 (2006)
16. Raghavan, U.N., Albert, R.K., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76, 036106 (2007)
17. Barber, M.J., Clark, J.W.: Detecting network communities by propagating labels under constraints. Physical Review E 80, 026129 (2009)
18. Liu, X., Murata, T.: An Efficient Algorithm for Optimizing Bipartite Modularity in Bipartite Networks. JACIII 14, 408–415 (2010)
19. Davis, A., Gardner, B.B., Gardner, M.R.: Deep south: A Social Anthropological Study of Caste and Class University of Chicago Press Chicago (1941)
20. Freeman, L.C.: Finding social groups: A meta-analysis of the southern women data. In: Dynamic Social Network Modeling and Analysis, pp. 39–97. National Academies Press (2003)
21. Newman, M.E.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98, 404–409 (2001)
22. Department of Finance and Deregulation. Dataset. Historical Australian Government Contract Data (February 27, 2013), `http://data.gov.au/dataset/historical-australian-government-contract-data/`
23. NSW Bureau of Crime Statistics and Research. Dataset. NSW Crime data (December 2008), `http://data.gov.au/dataset/nsw-crime-data/`