Network Science **8** (S1): S145–S163, 2020
doi:10.1017/nws.2020.11

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL ARTICLE

Is it correct to project and detect? How weighting unipartite projections influences community detection

Tristan J. B. Cann*, Iain S. Weaver and Hywel T. P. Williams

College of Engineering, Mathematics and Physical Sciences, Harrison Building, Streatham Campus, University of Exeter, North Park Road, Exeter, EX4 4QF, UK (e-mails: i.s.weaver@exeter.ac.uk, h.t.p.williams@exeter.ac.uk)
*Corresponding author. Email: tc471@exeter.ac.uk

Special Issue Editors: Hocine Cherifi, Stanley Wasserman

Abstract

Bipartite networks represent pairwise relationships between nodes belonging to two distinct classes. While established methods exist for analyzing unipartite networks, those for bipartite network analysis are somewhat obscure and relatively less developed. Community detection in such instances is frequently approached by first projecting the network onto a unipartite network, a method where edges between node classes are encoded as edges within one class. Here we test seven different projection schemes by assessing the performance of community detection on both: (i) a real-world dataset from social media and (ii) an ensemble of artificial networks with prescribed community structure. A number of performance and accuracy issues become apparent from the experimental findings, especially in the case of long-tailed degree distributions. Of the methods tested, the "hyperbolic" projection scheme alleviates most of these difficulties and is thus the most robust scheme of those tested. We conclude that any interpretation of community detection algorithm performance on projected networks must be done with care as certain network configurations require strong community preference for the bipartite structure to be reflected in the unipartite communities. Our results have implications for the analysis of detected community structure in projected unipartite networks.

Keywords: bipartite networks; unpartite projection; community detection; edge weighting

1. Introduction

Bipartite networks are a useful representation of many real-world systems where well-defined relationships exist between two distinct classes of nodes, such as scientific papers and their authors, or digital media and the people who share it. The complexity and relative obscurity of methods to analyze bipartite networks lead to frequent use of a unipartite projection of the system, so that more established unipartite methods can be applied (e.g. Alzahrani & Horadam, 2014; Del Vicario *et al.*, 2017; Newman, 2001). Another motivation for such analysis arises in situations where one class of nodes is used only to infer relationships between the other through projection (such as applying a network of users and reviews to identify groups of review spammers, Wang *et al.*, 2016). Unipartite projection encodes the edges between the two modes as a new network with only the nodes from a single mode, where nodes with a shared neighbor in the bipartite network are now directly connected. A cornerstone assumption is that the projected network retains key relationships such that community detection algorithms are able to capture structures which are meaningful in the bipartite context.

There are three main reasons for the use of unipartite projection in the study of bipartite networks. First, methods for directly analyzing bipartite networks are limited in their scalability and

their availability. Such methods, specifically designed to account for the additional complexities inherent in bipartite networks, are not widely included in popular network analysis packages and where such tools do exist they are not as capable of handling the large-scale datasets of modern network science. Taking the unipartite projection allows scientists to leverage the existing toolkits for unipartite networks. It is worth mentioning that it is possible to apply unipartite methods to bipartite networks by effectively discarding their bipartite structure. In the case of community detection, this approach is less accurate than using projection-based or bipartite methods (Arthur, 2019), but does present an alternative means of handling large bipartite networks. In addition, direct application to bipartite networks violates the assumption of edge independence in the definition of modularity (Newman, 2006). The remaining two points motivating the use of unipartite projection are best framed within use cases. In many experimental settings, one of the two bipartite modes is the primary focus. For example, given a network of authors and publications, we may study coauthorship using publications solely to infer edges between authors. Hence, projecting the network focuses on the specific area of interest. This ties closely into the final reason to consider how unipartite projection affects community structure—some unipartite networks are implicitly projections of some hidden bipartite network. Consider again the coauthorship network. At first glance this is a unipartite network, but it is in fact an implicit projection of a bipartite network between scholars and the institutions and events they have visited—it is very unlikely for coauthorship to arise without such a meeting.

In our previous work, we studied the efficacy of bipartite community detection using unipartite projections (Cann *et al.*, 2019), constructing an ensemble of synthetic bipartite networks with imposed community structure and attempting to recover the structure from unipartite projections made with four candidate projection schemes. Here we extend our previous assessment by including three additional unipartite projection schemes. We also present a comparison of communities found using the seven projection methods applied to a real-world dataset, in this case a bipartite network linking web pages (URLs) to the Twitter users who shared them during one week of conversations about climate change.

Several community detection methods have been shown to be effective at partitioning small bipartite networks; Barber (2007) adapts the null model used to compute modularity (Newman, 2006) on unipartite networks to the bipartite case to account for the additional requirement that the vertices incident to each edge must be in different modes. Beckett (2016) reports other approaches, including weighted bipartite modularity maximization. However, optimization of bipartite modularity (e.g. through implementations such as the MODULAR package (Marquitti *et al.*, 2014) or those reported by Beckett (2016)) is computationally demanding and may be of limited use on large networks, such as those from online social media. Beyond algorithms using a modularity maximization approach, efforts have been made to extend stochastic block model (SBM) methods to bipartite networks. Larremore *et al.* (2014) and more recently Wyse *et al.* (2017) have demonstrated that enforcing the two node types required of a bipartite network allows the discovery of meaningful community structure. Many of these bipartite community detection methods find clusters which contain only a single node type and as such there are no edges within each community. This behavior is advantageous since it allows different numbers of communities to be found in each mode with many-to-one correspondences between them, but counterintuitive from the perspective of unipartite communities. Despite the quality improvements achieved using bipartite community detection methods, they are less widely implemented in network analysis packages, furthering the appeal of unipartite projections.

Bipartite networks are an intuitive representation of social media activity, such as where Del Vicario *et al.* (2017) examine how Facebook users interact with information related to the 2016 EU Referendum in the UK as two network modes. By computing the unipartite projection onto page nodes, they identify communities of pages within which groups of users more frequently interact. Schmidt *et al.* (2017) use a similar methodology to identify and explore the user groups formed around frequent likes or comments on the same Facebook content. Twitter is another

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

platform readily studied using the unipartite projection approach. Williams *et al.* (2016) explore behavior patterns amongst Twitter users and the news articles they share through a projection onto the article network. Analysis of this projected network found communities of news domains that were frequently shared by the same users.

Such analysis is also suited to physically embedded networks such as where Chen *et al.* (2007) study one representation of the Chinese bus transport network as the projection onto both modes of a stop-route network. Srivastava *et al.* (2013) apply projection to bipartite networks of documents and terms to find clusters of similar documents. They use a threshold approach for the unipartite edges, discarding those that have a weight lower than a fixed value. Alzahrani & Horadam (2014) apply unweighted projections to two crime-related networks, finding a topographical division between urban and rural municipalities when looking at crime in New South Wales, Australia, and communities encompassing training links in a terrorist-activity network. Isah *et al.* (2015) study the network of people and crimes to find different types of organisational structures among perpetrators. Yan & Ding (2012) study various networks arising from authorship and citation behaviors and assess the similarity between topic, coauthorship, and citation networks.

These are all examples where a bipartite network is explicitly projected, but numerous other studies encode this process in the network construction such as Starbird (2017) who studies alternative news domains shared by Twitter users around mass shooting events. Newman (2006) considers book co-purchases when testing a spectral method for community detection, which is an implicit projection of the user-item network. The design choices implicit in construction of these unipartite networks are subject to the same biases and pitfalls inherent in projection schemes.

Although unipartite projection and community detection see frequent use in empirical work, limited theoretical study has been devoted to how the community structure in a projected network relates to the community structure in the original bipartite network. Everett & Borgatti (2013) show that while unipartite projection onto a given mode results in a loss of information (since encoding one class of nodes as edges between the other is generally not reversible), it is possible to derive meaningful results by considering projections onto each mode simultaneously. Melamed (2014) takes the concept of dual-projection to refine the community detection process on the bipartite networks by incorporating information from both unipartite projections. Arthur (2019) extends this through a comparison of modularity metrics by including a novel modularity formulation that accounts for structure in the bipartite network. As noted by Newman (2001), high-degree nodes in a bipartite network contribute a disproportionate number of edges to the corresponding unipartite projection; a node of degree k contributes of the order k^2 edges to the projection. Certainly, we must be careful when weighting these edges. Guimerà *et al.* (2007) define a model to generate bipartite networks with a fixed community structure, where the parameter p denotes the fraction of network edges which join nodes within prescribed communities. They also adapt the standard definition of modularity to better reflect bipartite network structure, and test this against weighted and unweighted projections. The key finding is that in some cases both unipartite and bipartite modularity have similar performance. Bongiorno *et al.* (2017) incorporate statistically validated networks into the community detection process by finding stable cores within bipartite communities. Li & You (2013) examine whether any network metrics are affected by the unipartite projection process and find that certain metrics such as clustering coefficient vary with projection scheme, while degree correlation does not.

A common first step in constructing a unipartite projection is to filter out low-weight edges such as by establishing a threshold and removing those that do not meet a specified criteria, a step which can be very helpful computationally by dramatically reducing the edge density in the projection. Sasahara (2016) constructs word association networks by calculating cosine similarity between word contexts and retaining only those edges that exceed a given weighting threshold. Other methods compute edge significance relative to a null model to determine which edges to keep. Grinberg *et al.* (2019) find networks of news sources which are visible to the same people

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

on social media through a multiscale backbone approach. Saracco *et al.* (2017) make use of exponential random graph models to determine statistical significance of the edges in the unipartite projection, and retain only those edges that satisfy a given significance threshold. Thresholding methods can have merit in certain use cases, but as with bipartite community detection methods, they are often not included in the most widely used libraries. Another potential issue with how these methods have been used in the past is the binarizing of the remaining edge weights. While the use of thresholding is likely to increase accuracy over the case of a fully binarized projection, the loss of information in the significant edge weights is not to be overlooked. The final concern with removing edges given some thresholding criteria is fragmentation in the network. Many applications of bipartite network projection consider only the giant component, and if sufficient fragmentation occurs, it is likely that the size of the giant component will no longer be comparable to the size of the whole network.

The methods described so far have found particularly strong purchase in the study of social systems such as online social media and scientific collaboration. Along with many natural systems, the networks in such studies often have degree distributions with a long tail, where a small number of profoundly well-connected nodes exist in a sea of low-degree nodes. The popularity and success of this approach on these systems suggests that these properties are beneficial to established methods, although careful consideration needs to be given to how the properties of the bipartite degree distribution influence the structure of the projected unipartite network and the performance of community detection methods.

In this study, we consider a range of network and projection types under community detection, and evaluate the quality of the output against prescribed communities. Networks are differentiated by their degree distributions, selected to include those characterized by geometric-, binomial-, and zeta-like tails and the spread of edges within and between prescribed communities. Unipartite projections are taken using seven different edge weighting schemes, before testing the ability of unipartite community detection algorithms to recover bipartite community structure. We first illustrate the different outcomes associated with each projection scheme applied to a real-world bipartite network. We next perform a more rigorous test that seeks to recover bipartite community structure after unipartite projection of a series of synthetic networks with imposed community preference. In Section 2, we detail the real-world dataset studied, the network model used to sample synthetic networks, the seven projection weighting schemes used, and the metrics by which we measure community detection performance. Section 3 presents the results of our experiments, and Section 4 discusses their consequences.

2. Methods

This section begins by explaining the methodology to construct and study a network from a Twitter dataset. Also in this section, we outline a model for generating random bipartite networks with a prescribed community structure and distinct degree distributions. Seven different weighting schemes are described for use in unipartite projections, and finally we outline the process for computing and evaluating the accuracy of community detection on these network projections. We use *left* and *right* as generic labels for our bipartite modes throughout this paper. The results from these methods are presented in Section 3.

2.1 Real-world dataset

For our study of how the different projection weighting schemes affect real-world networks, we make use of a Twitter dataset to construct a bipartite network between users and the URLs they share. The dataset was gathered from the Twitter Streaming API[1] using the search terms *climate change* and *global warming* for a one week period between May 31 and June 06, 2017. We keep only those tweets containing URLs which can be resolved and remove any user that shared URLs

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

more than 50 times during the week as a likely automated account such as news aggregators. Finally, we apply a disambiguation step to URLs by following any permanent redirects to reveal the destination of masked URLs, such as those from link shortening services. This leaves a dataset of 187, 378 tweets by 54, 347 users sharing 20, 880 distinct URLs.

From this dataset, we construct a bipartite network linking user nodes to URL nodes, adding an edge whenever the user includes the URL in one of their tweets. Edge weights are assigned as the number of times a user shared the same URL. This gives a bipartite network with 80, 009 edges (numbers of user and URL nodes as above). We restrict to the giant component for all further analyses, which contains 7, 496 URL nodes, 42, 113 user nodes, and 63, 755 edges. Taking the projection onto the URL nodes of the giant component gives a unipartite network with 53, 652 edges.

2.2 Synthetic bipartite networks

2.2.1 Generative bipartite network model

Many different methods can be developed for producing synthetic bipartite networks; free variation of network statistics such as edge density, degree distribution, and vertex correlation may result in a wide range of structures and behaviors. Our model was designed to minimize the number of assumptions made by prescribing only the degree distribution and the vertex correlations required to impose community structure on the network. Even in this case of limited assumptions, there are many candidate degree distributions to choose from. Here we construct a generative model motivated by a simple physical interpretation, described in detail by Weaver (2015).

If we consider a growth process, we add new vertices at a constant rate, with a number of incident edges governed by preferential attachment. A parameter m determines the level of preference for high-degree nodes in the assignment of new edges. We study two cases: $m \to \infty$ (no preferential attachment) and finite m (a strong preference for high-degree vertices). Without preferential attachment, the growth process still produces an interesting degree distribution. The main difference between the two cases is that preferential attachment yields a zeta distribution (or colloquially a "power law") as vertex degree k increases in the tail of the distribution. Preferential attachment therefore leads to the formation of extremely well-connected vertices, a feature typical of self-organizing structures in nature and human society (Newman, 2005). Real-world phenomena, such as the distribution of page interactions on Facebook (Del Vicario *et al.*, 2017), are well represented by this model. With no preferential attachment, a geometric degree distribution emerges. These two cases for m enable a comparison of how heavy-tailed distributions of different types interact with the projection process.

Previous work by Weaver (2015) derived the steady-state degree distribution of randomly grown networks under different preferential attachment conditions as:

$$\mathrm{pmf}(k) = \frac{m+\delta}{m(\delta+1)+1} \frac{(m)_{(\frac{m}{\delta}+2)}}{(m+k)_{(\frac{m}{\delta}+2)}} \tag{1}$$

making use of the Pochhammer notation $(a)_b = a(a+1)...(a+b-1)$. The parameter δ defines the number of edges incident to each newly added vertex. In the case $\delta = m = 4$, this simplifies (for sufficiently large k) to $\mathrm{pmf}_{zeta}(k) \sim k^{-3}$, that is, a zeta distribution representing preferential attachment. Setting $m \to \infty$ and $\delta = 4$ gives $\mathrm{pmf}_{geo}(k) = \frac{1}{5}\left(\frac{4}{5}\right)^k$, that is, a geometric distribution from growth without preferential attachment. In contrast to these physically motivated models, we provide an Erdös–Renyi bipartite graph for comparison, with degree distribution given by $\mathrm{pmf}_{bin}(k) = \binom{\delta N}{k} N^{-k}\left(1-\frac{1}{N}\right)^{\delta N-k}$. Figure 1 plots each of these degree distributions.

After choosing a degree distribution, construction of a random network begins by assigning the number of vertices in the left and right modes, $N_l = N_r = N$, respectively, and sampling degrees from the distribution for each vertex. In all cases, both bipartite modes sample from the same

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

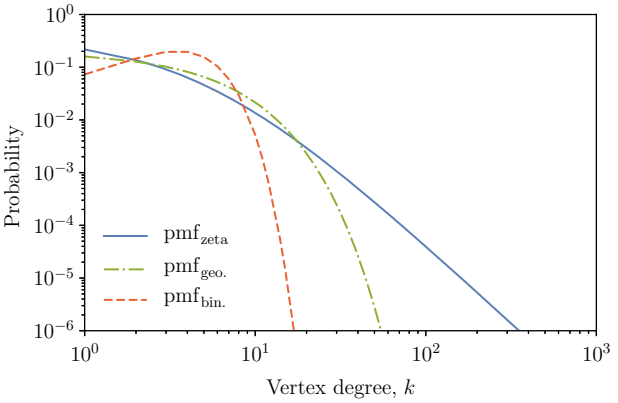

Figure 1. The three degree distributions used in the generative model. Each distribution has identical mean $\langle k \rangle = 4$, but vary in the weight of the tail as k increases.

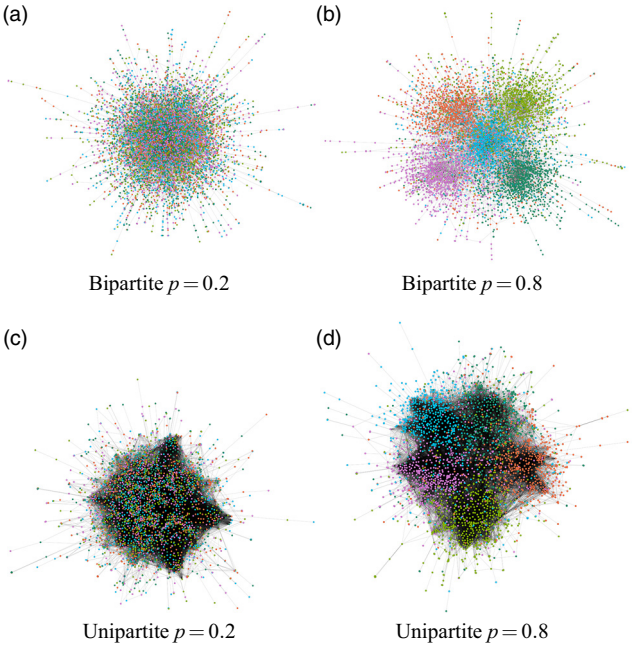

Figure 2. The giant components of bipartite networks produced by a single network instance of 10^4 nodes with average degree 2 divided into $M = 5$ communities along with their unipartite projections: (a), (c) have low community preference $p = 0.2$, while (b), (d) use high community preference, $p = 0.8$. Node colors indicate community assignment. The dramatic increase in edge density across the projection process can be seen by the relative intensity of black edges in (c), (d). Networks are visualized using Gephi with layout determined by the ForceAtlas2 algorithm (Jacomy *et al.*, 2014).

degree distribution. Edge creation is performed by randomly selecting pairs of nodes, choosing one from each mode weighted by their unassigned degree. We impose community structure following the method of Guimerà *et al.* (2007) by defining a partition of the vertices into M equally sized communities before assigning edges, with a one-to-one correspondence between the communities in each mode. We define a parameter p to fix the probability of an edge connecting two vertices in the same community, with complementary probability $1 - p$ of connecting vertices regardless of their assigned communities. Notably, the proportion of edges joining vertices in the same community is not simply p, but $p + \frac{(1-p)}{M}$, which varies from $\frac{1}{M} \to 1$ as p varies over $0 \to 1$. Many vertices have degree $k = 0$, a characteristic frequently mirrored in real-world citation networks (Larivière *et al.*, 2009). We discard isolated nodes before continuing our analysis. Sample network giant components produced by this model, and their projections, can be seen in Figure 2.

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

2.2.2 Producing an ensemble of synthetic networks

In order to study the expected behaviour of different projection weighting schemes under community detection, we construct an ensemble of synthetic networks with known community structure as outlined in Section 2.2.1. Each network consisted of $N = 10^6$ nodes (in each mode) divided into $M = 5$ communities. For each degree distribution, we fix the expected node degree to be 4, giving approximately 4×10^6 edges in each synthetic network. For each value of the community preference parameter $p \in \{0, 0.1, 0.2, \ldots, 1\}$, networks are generated for each degree distribution outlined by Figure 1. We present our results as averages over 100 network realizations for each combination of projection weighting, degree distribution, and community preference p.

2.3 Unipartite projection and edge-weighting schemes

Taking the unipartite projection produces an edge between each node pair of the chosen class with at least one shared neighbor in the other class. A key part of this process is how edge weights are calculated in the projection; approaches can be as simple as recording presence of a mutual neighbor or as complex as nonlinear weighting from the overlap of the neighborhood sets. Each of the weightings outlined below has been designed to work with both weighted and unweighted bipartite networks. In practice, however, the bipartite networks tested (both empirical and synthetic) have primarily binary edges, that is, the proportion of edges with weights larger than one is small. Here we detail seven weighting schemes for edges in the unipartite network which will be tested against the quality of unipartite community detection relative to different levels of bipartite network structure. As outlined here, the methods describe projection onto the right nodes, but apply equally to the left nodes under suitable transposition. Given this flexibility in application, the choice of which projection to use is circumstantial and depends on the research question. In our case, the choice is arbitrary; the generative model produces networks with statistically symmetric modes and hence statistically symmetric projections. In an experimental setting, it is often clear which of the two modes is of interest, making it obvious which nodes should be projected onto.

The *simple* weighting scheme calculates the weight w_{ij} for edge e_{ij} in the projected network as the number of neighbors nodes i and j share in an unweighted bipartite network. In the case of a weighted network, w_{ij} represents the sum of the product of edge weights on all i, j-paths of length two. Under simple weighting, the bipartite adjacency matrix B (with rows corresponding to the left-mode, columns corresponding to the right-mode, and entries corresponding to the edge weights) is used to define the unipartite adjacency matrix:

$$\mathrm{U}_{\mathrm{simple}} = \mathrm{B}^{\top}\mathrm{B}. \tag{2}$$

Note that U is symmetric, and has nonzero diagonal elements and as such encodes a network with self-connections. In this paper, we do not allow self-connections, and set the diagonal elements to zero.

The *binary* weighting scheme is calculated from the simple edge weights by truncating at 1; that is, we do not consider the number of shared neighbors between a pair of nodes. As such, we record the presence or absence of a shared neighbor as a 1 or 0, respectively, giving:

$$\mathrm{U}_{\mathrm{binary}}[i,j] = \begin{cases} 1 \text{ if } \mathrm{U}_{\mathrm{simple}}[i,j] \neq 0, \\ 0 \text{ if } \mathrm{U}_{\mathrm{simple}}[i,j] = 0. \end{cases} \tag{3}$$

The *hyperbolic* weighting scheme, introduced by Newman (2001), is a means to limit the influence of high-degree nodes in the bipartite network in the projected network. A node of degree k in the bipartite network will contribute a total edge weight proportional to the square, $\frac{1}{2}k(k-1)$ under the simple weighting scheme. As a result, high-k nodes can have a disproportionate influence on total edge weight and consequently community quality in the projected network. This is

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

of particular concern in networks with long-tailed degree distributions. The hyperbolic scheme applies a scaling factor of $(k_i - 1)^{-1}$ to each edge created in the projection of node i with degree k_i. In this scheme, high-degree nodes still have an increasing contribution to the total edge weight, but now contribute linearly by degree, $\frac{1}{2}k$. Under hyperbolic weighting, the unipartite adjacency matrix is defined as:

$$\mathrm{U}_{\mathrm{hyper.}} = \mathrm{B}^{\top}\mathrm{W}\mathrm{B} \text{ where } w_{ij} = \begin{cases} (k_i - 1)^{-1} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The *unary* weighting scheme extends Newman's hyperbolic weighting to normalize each node's contribution to the total edge weight in the projected network. The edge weights formed by projection of node i are rescaled by $2k_i(k_i - 1)^{-1}$. As a result, the total edge weight contribution of a node to the projected network is exactly 1. Under unary weighting, the unipartite adjacency matrix is defined as:

$$\mathrm{U}_{\mathrm{unary}} = \mathrm{B}^{\top}\mathrm{V}\mathrm{B} \text{ where } v_{ij} = \begin{cases} 2(k_i(k_i - 1))^{-1} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The *random walk* weighting scheme evaluates the probability distribution of two-step random walks on the bipartite networks to determine the projected edge weights. Such an approach is often used when calculating node ranking or similarity for recommender systems (e.g. Lee *et al.* (2011)). This is calculated by row-normalizing the bipartite adjacency matrix and performing a matrix multiplication as with previous methods, that is,

$$\mathrm{U}_{\mathrm{randw}} = |\mathrm{B}^{\top}|_{\mathrm{L1}}|\mathrm{B}|_{\mathrm{L1}}, \tag{6}$$

where $|\mathrm{B}|_{\mathrm{L1}}$ denotes the matrix B after L1 normalization of the rows.

The *cosine* weighting scheme is a nonlinear measure of similarity between node neighborhoods. We define the weight of an edge between two nodes in the unipartite network as the cosine similarity of the two corresponding neighborhoods, that is,

$$\mathrm{U}_{\mathrm{cosine}}[i, j] = \frac{\mathrm{B}[:, i] \cdot \mathrm{B}[:, j]}{|\mathrm{B}[:, i]||\mathrm{B}[:, j]|}. \tag{7}$$

The *Jaccard* weighting scheme measures the overlap between nodes' neighborhoods. Bipartite edge weights can be incorporated by weighting neighborhood elements. The edge weight is defined as the ratio between the sizes of the intersection and the union of the node neighborhoods, that is,

$$\mathrm{U}_{\mathrm{Jaccard}}[i, j] = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}, \tag{8}$$

where $N(i)$ is the immediate neighborhood of node i.

We illustrate the effects of each of these weighting schemes on community detection in the unipartite projections in Figure 3 by constructing a small, unweighted bipartite network such that the community structure found on each projection is unique.

2.4 Community detection

We produce bipartite networks by varying the parameter p, which controls the preference of nodes to connect with other nodes within their prescribed communities. Given a unipartite projection, we apply community detection with the expectation that we can recover some amount of the community structure used in construction. When analyzing the networks produced by the generative model, we measure the accuracy in the returned community partition with respect to the prescribed communities.

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

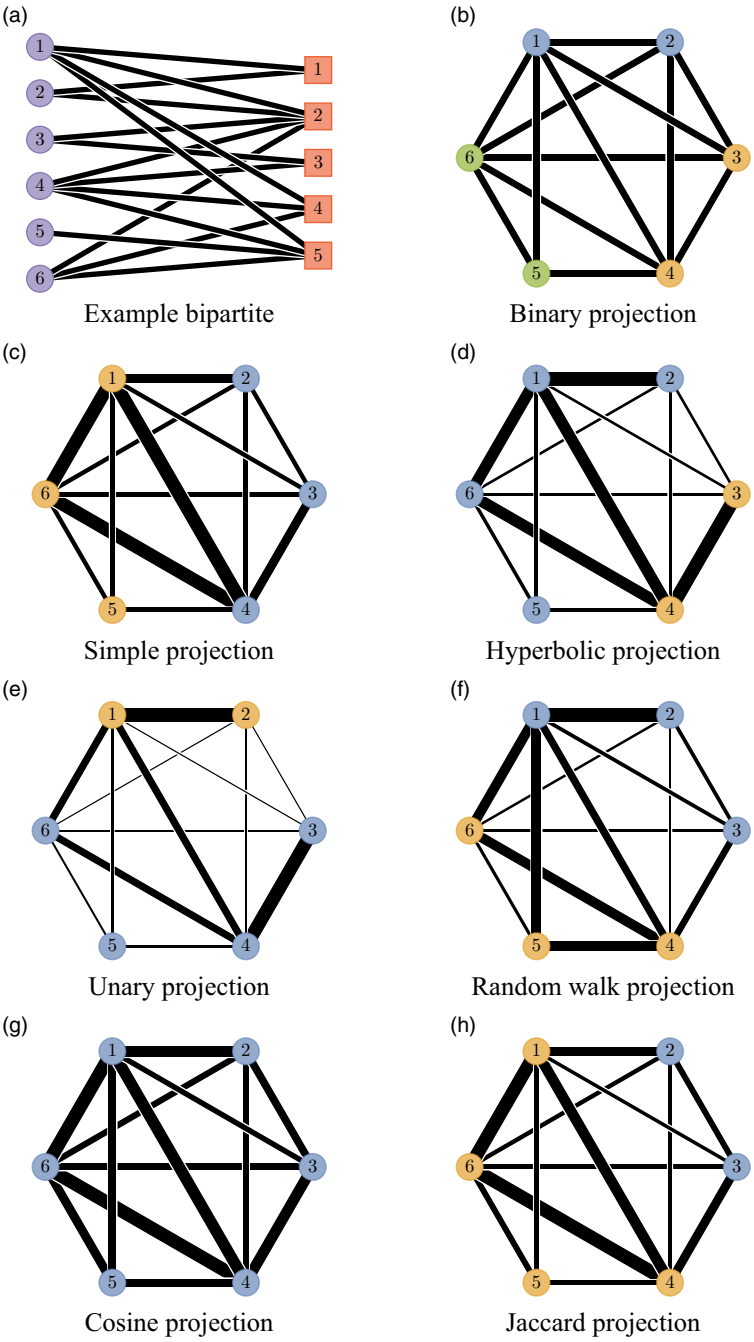

Figure 3. A small, unweighted bipartite network and its unipartite projections. The node color in the bipartite denotes the left and right modes, the node colors in the unipartite projections denote the communities found by the Louvain algorithm (Blondel *et al.*, 2008), and line thickness is proportional to edge weight. The bipartite network was constructed to ensure that the detected community structure is different under each projection to highlight the impact of projection weighting on community detection.

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

In all cases, we use the Louvain algorithm proposed by Blondel *et al.* (2008) for community detection. This algorithm estimates the best community partition through modularity maximization on the large and locally dense networks produced by our model. The algorithm begins by assigning each node to its own community, iteratively merging neighboring communities which produce the largest increase in network modularity, with ties broken by random selection. When no more steps can increase modularity, a new network is induced by merging all nodes in a community into a single node, and the first step is repeated on the new induced graph. This method proves highly scalable, allowing calculation of communities in large, weighted networks. In our case, we consider unipartite modularity, but by changing the modularity function, the algorithm can be applied to other network types. The Louvain algorithm requires no information about the number of communities to find. This behaviour is ideal for many experimental use cases as it means the final partition is decided entirely by network topology.

2.5 Assessing the accuracy of community detection

Our detected network communities are evaluated against the prescribed community labels by using the adjusted Rand index, that is, the proportion of all node pairs for which community labels are either the same or different in *both* the computed and prescribed labeling, adjusted by the expected level of agreement by chance:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}(\mathrm{RI})}{\max(\mathrm{RI}) - \mathbb{E}(\mathrm{RI})}. \tag{9}$$

The adjusted Rand index takes values in $[0, 1]$, where values close to 0 indicate that agreement between the true and detected communities is no better than chance, and 1 indicates that the true and detected communities are identical. This measure has the desirable property that the precise community labels found are not important for evaluation of the adjusted Rand index, only whether two given nodes have the same community assignment in the detected and reference structures. As a result, permutation of the community labels does not affect this measure. We choose the adjusted Rand index over other information theoretic measures for two reasons. We argue that it is important for any comparison of our community structures to account for chance agreement. The adjusted Rand index explicitly accounts for this using a null model, whereas competing measures such as normalized mutual information do not. We also follow the advice of Romano *et al.* (2016) who find that the adjusted Rand index performs better when considering relatively few different labels in the reference partition.

We also compare the sizes of the detected communities to the sizes known in our synthetic networks. We do this by computing the expected community size of a random node, that is,

$$\frac{\sum_C |C|^2}{\left(\sum_C |C|\right)^2}. \tag{10}$$

where $|C|$ denotes the number of nodes in community C. Note that this measure is distinct from the mean community size which would be heavily skewed by a large number of small communities, a typical result of community detection on any large network.

3. Results

3.1 Example: Real-world network from Twitter

We first report on the communities found in unipartite projections of the Twitter network made using the different weighting schemes. Figure 4 visualizes the communities found in the seven projections of the giant component, with a layout determined by the ForceAtlas2 algorithm (Jacomy *et al.*, 2014). Node colors depict community membership sorted by community size.

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

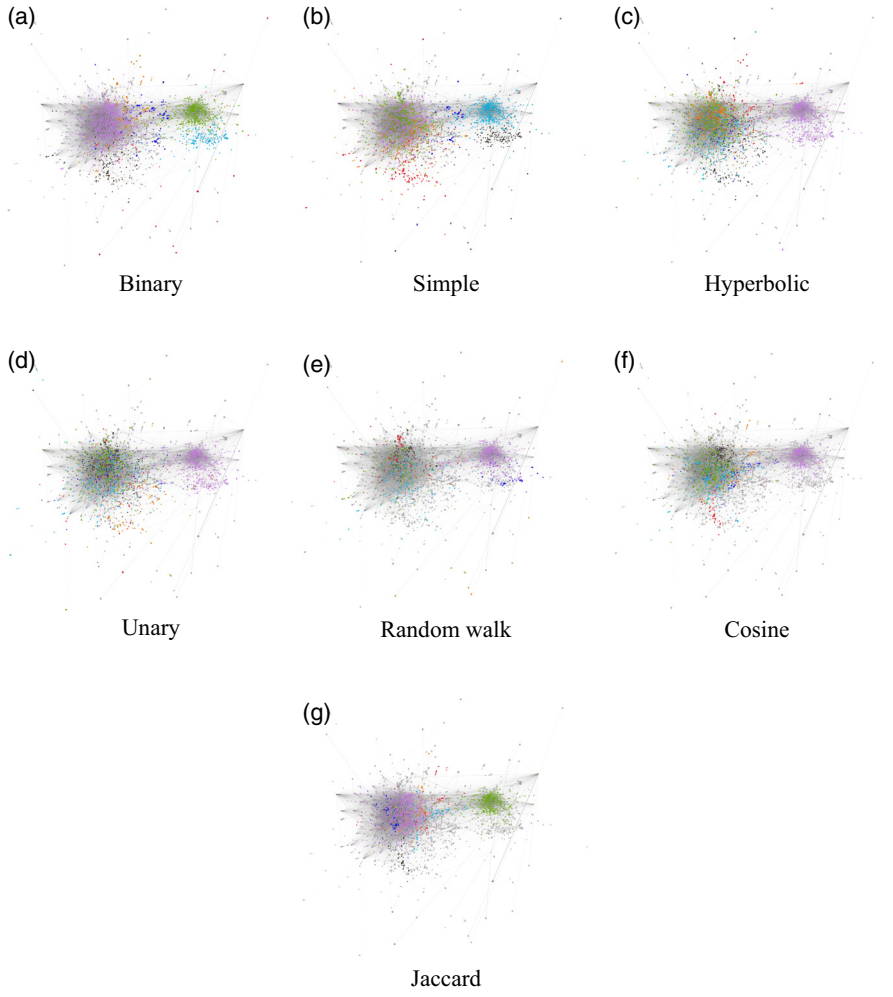

Figure 4. Community structure found when applying the seven different weighting schemes to the projection of the Twitter network. Only nodes of degree at least 5 are visible, and node color corresponds to communities in decreasing size order (pink, green, light blue, black, orange, red, blue, grays, respectively). Note the variability of the division and size ranking of different communities under each of the seven different weighting schemes. For references to color, the reader is directed to the online version of this article.

Three types of community assignments appear across the seven projection weighting schemes. In the first type, the largest community dominates the left-hand cluster of the network (as seen in Figures 4(a) and (g)). In the second type, the largest community dominates the right-hand cluster (as seen in Figures 4(c), (d), and (f)). In the final type, each cluster is made up of multiple smaller communities.

As we do not know the "true" community structure for the empirical Twitter network, we cannot compute the accuracy of community detection. Instead, we can compare the community partitions found by different projection schemes with each other. Table 1 shows the expected community size, the sizes (number of nodes) of the 5 largest communities detected, and the corresponding modularity. The random walk, Cosine, and Jaccard methods stand out as producing extremely high modularity scores along with a large number of small communities. This seems counter intuitive and as we will see in Section 3.2, a high modularity in a projected network is not always a sign of underlying community structure. The binary weighting finds the single largest

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

Table 1. Statistics for the communities found by the Louvain algorithm (Blondel *et al.*, 2008) on the unipartite projection of the Twitter dataset under different weighting schemes.

Weighting	Expected community size	Size of 5 largest communities					Modularity
Binary	0.122	2,124	1,081	525	421	375	0.565
Simple	0.105	1,377	1,069	1,046	834	720	0.614
Hyperbolic	0.110	1,808	1,156	781	530	527	0.580
Unary	0.087	1,664	899	632	494	420	0.609
Random walk	0.032	829	451	333	248	238	0.791
Cosine	0.042	879	833	488	402	231	0.870
Jaccard	0.068	1,631	890	164	157	153	0.932

Table 2. Pairwise adjusted Rand index comparisons of the different community structures detected on projections of the Twitter network under the seven weighting schemes.

Weighting	Binary	Simple	Hyperbolic	Unary	Random walk	Cosine	Jaccard
Binary	1	0.401	0.285	0.23	0.191	0.279	0.416
Simple	0.401	1	0.368	0.292	0.21	0.26	0.23
Hyperbolic	0.285	0.368	1	0.621	0.205	0.227	0.197
Unary	0.23	0.292	0.621	1	0.207	0.214	0.173
Random walk	0.191	0.21	0.205	0.207	1	0.327	0.196
Cosine	0.279	0.26	0.227	0.214	0.327	1	0.408
Jaccard	0.416	0.23	0.197	0.173	0.196	0.408	1

community, but leads to a large number of smaller communities compared to the simple, hyperbolic, and unary weightings which perform similarly, with a broader distribution of community sizes. Calculation of the Gini coefficient on the distributions of community sizes over the different projection weightings finds that most of the detected partitions have high size inequality ($0.75 < G < 0.9$). The exception to this is the random walk weighting scheme which had a Gini coefficient of 0.409, indicating a very broad distribution of small communities.

Table 2 shows the pairwise adjusted Rand index between the community structures detected under each projection scheme. The general trend between the community assignments for nodes shows limited similarity under the different weighting schemes. A notable exception is the hyperbolic and unary schemes which are the most similar pair. Some similarity is also observed between the binary and simple weightings and the cosine and Jaccard weightings. Also of note is the random walk weighting, which gives the lowest average similarity to other methods.

Taken together, the results from applying different unipartite projection schemes to a real-world bipartite network give a good indication that the method of projection has a large impact on the community partition that is found. We do not know the true partition of this empirical network, so we cannot determine the accuracy of community detection by each method. However, the variations between outcomes for different methods raise the question of which projection method permits the most accurate identification of community structure.

3.2 Testing with synthetic network ensembles

In this section, we report results from a systematic exploration of community detection accuracy using unipartite projections of bipartite networks with known community structure. Since the two modes in our synthetic networks are generated and connected using the same processes, the left and right projections are statistically indistinguishable. As such, we report only results on the right projection. The Louvain community detection algorithm is applied to each projected network, and

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

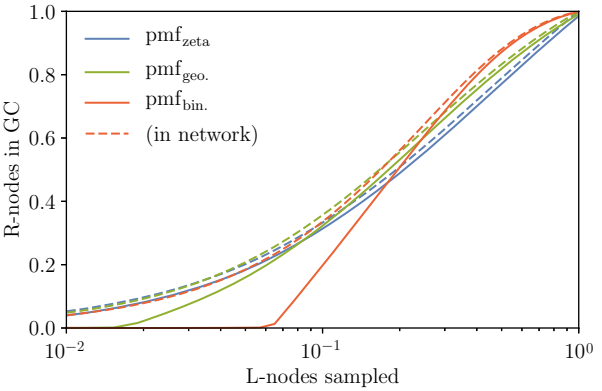

Figure 5. Fraction of all R-nodes included in the sampled network and its giant component for a given sample of L-nodes. These results are derived from network instances with 10^6 vertices in each of the L- and R-modes, 4×10^6 edges, and community preference $p = 0.5$.

we evaluate the accuracy of the resulting partition using the adjusted Rand index, modularity, and expected community size.

A key consideration when constructing networks from real-world datasets is sampling; how complete must a sample from one set of nodes be in order to recover a dataset representative of the system as a whole? To give a more concrete example, suppose that we wish to sample the authorship network by taking all works by a number of scholars. How many authors are we required to sample to produce a network that has a giant component of the necessary size? By constructing models of bipartite networks with different degree distributions, we can provide insights for a breadth of relevant networks. We iteratively sample nodes from one mode, in this case right, computing what proportion of nodes in the left mode we discover and furthermore what proportion are connected to the largest network component. The results displayed in Figure 5 show that for a geometric-tailed degree distribution, and more so for the binomial distribution, we see a similar effect to Callaway *et al.* (2001), where a finite sample is required to produce a sample network with a giant component. In contrast, networks with long-tailed degree distributions have a giant component which can be recovered from very small vertex samples as a consequence of the high-degree "hub" nodes; a phenomena which may be credited in part for the success and growth of this field.

As previously mentioned in establishing the network size, an additional computational challenge exists around projecting networks with heavy-tailed degree distributions. Recall that a bipartite vertex with degree k produces $\frac{1}{2}k(k-1)$ unipartite edges after projection. This typically leads to a dramatic increase in the edge density of the projected network. With 10^6 nodes in each mode and 4×10^6 edges, the projected network from binomial, geometric, and zeta distributions result in roughly 8×10^6, 16×10^6, and 120×10^6 edges, respectively; that is to say the long-tailed degree distribution experiences a 15-fold increase in network density over the binomial degree distribution. The number of edges in a unipartite projection is tied to the second moment of the bipartite degree distribution; long-tailed distributions frequently have divergent second moments (as in this case), which cause the number of edges to grow rapidly with the size of the network, and produce dense unipartite projections.

Figure 6 reports three different metrics for the performance of community detection on the unipartite projections: adjusted Rand index, unipartite modularity, and expected community size for a uniformly chosen node. The extent to which prescribed communities can be recovered computationally is strongly dependant on all of our model parameters; the node degree distribution, the strength of imposed community structure, and projection weighting. In particular, agreement with prescribed community labels is only found with strong imposed community structure at high p.

The adjusted Rand index results in Figure 6(a) show the extent of agreement between the prescribed and detected community labels. A near-zero value indicates that community labels are

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

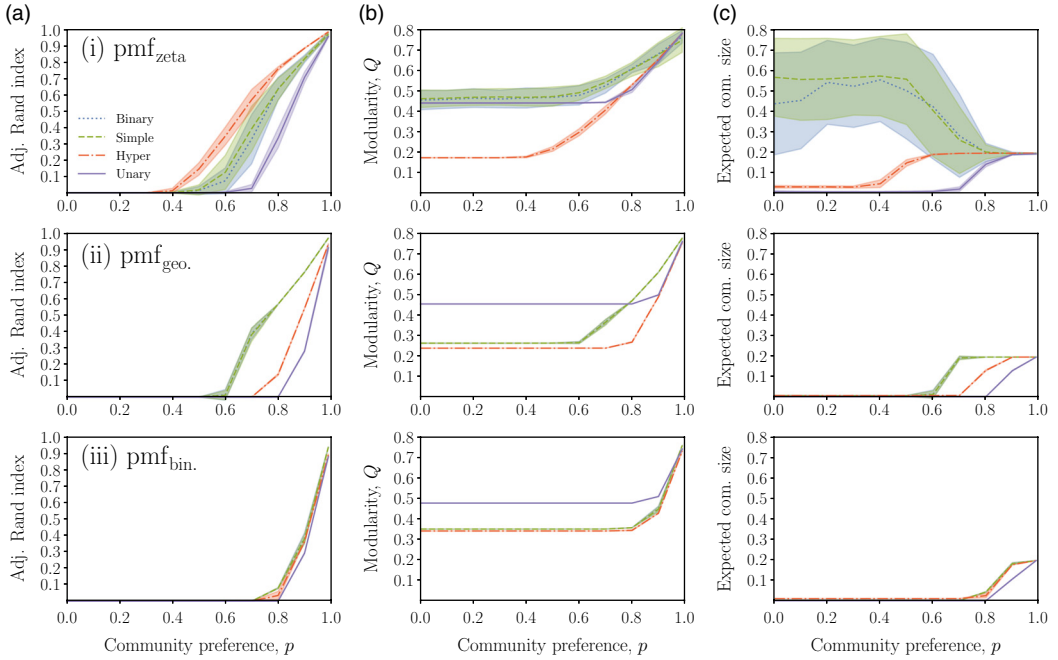

Figure 6. Comparison of community detection after binary, simple, hyperbolic, and unary weighted projections. Lines show the mean over 100 iterations, and the shaded region indicates ± one standard deviation. Left to right: (a) Adjusted Rand index, (b) modularity, and (c) expected community size, across bipartite networks with varying levels of community structure. Top to bottom: (i) pmf$_{zeta}$, (ii) pmf$_{geo.}$, and (iii) pmf$_{bin.}$ bipartite degree distributions.

a no-better predictor of true values than a random assignment, while increasing values indicate better performance. Figure 6(a) shows that for a long-tailed degree distribution, community detection reveals meaningful community labels at a much lower threshold of community preference, approximately $p \geq 0.4$ compared to values of roughly 0.6 and 0.7 for geometric and binomial degree distributions, respectively. In all cases, the weighting scheme has a significant impact on performance, with hyperbolic weighting outperforming other schemes, and unary weighting performing much worse.

The modularity (Figure 6(b)) and expected community size (Figure 6(c)) results provide additional insight into the recovery of the underlying community preference by community detection on the unipartite projection. The cause appears to be that modularity in the unipartite projection is not completely determined by the level of imposed community structure, as the expected increasing behavior only occurs with large p. Furthermore, modularity is shown to be nonzero for partitions on networks with weak or no imposed community structure (as high as 0.5). This suggests high-quality partitions have been identified in the projected network, although in the generative method, we have imposed weak or no bias at all. The hyperbolic weighting scheme demonstrates a desirable effect in this regard, returning the lowest modularity for low p, whereas the unary weighting scheme performs worst, giving high modularity for low p. In the hyperbolic and unary cases, we find that the expected size of community is small when the adjusted Rand index indicates poor recovery of the imposed community structure, before converging to 0.2, the value of the true partition. Exceptions are the binary and simple weightings which decrease to the convergent value. Combined with the modularity results, this suggests that at low p the modularity maximizing algorithms find high modularity by partitioning the network into one or more large communities. More meaningful communities emerge with increasing p. This behavior is likely caused by the dominance of large cliques formed in the projection of high-degree nodes.

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

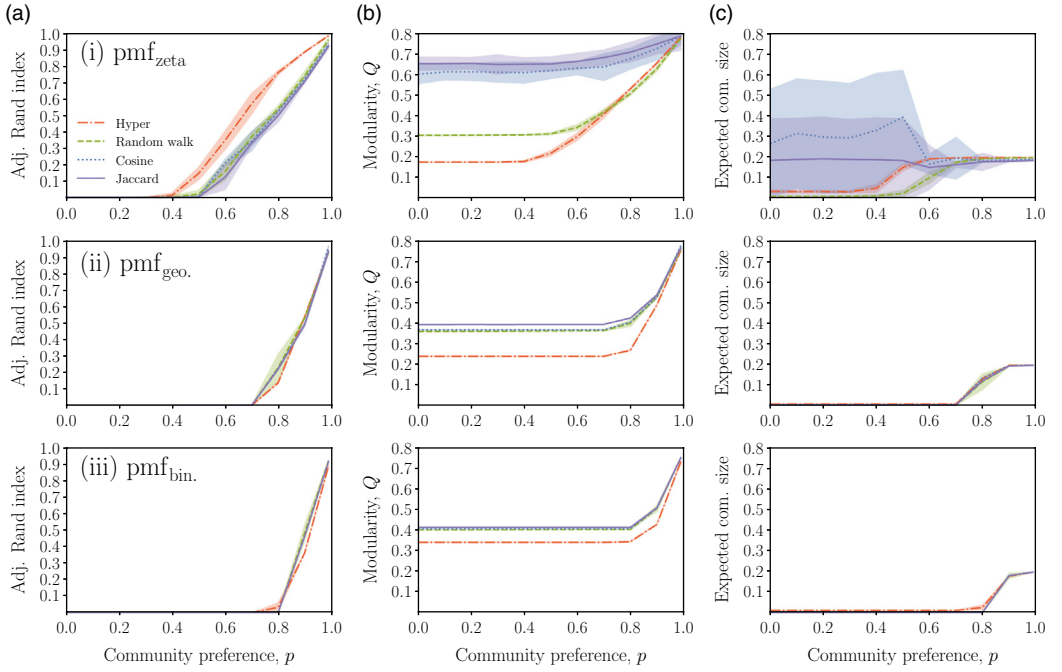

Figure 7. Comparison of community detection after hyperbolic, random walk, cosine, and Jaccard weighted projections. Lines show the mean over 100 iterations, and the shaded region indicates ± one standard deviation. Left to right: (a) Adjusted Rand index, (b) modularity, and (c) expected community size, across bipartite networks with varying levels of community structure. Top to bottom: (i) pmf$_{zeta}$, (ii) pmf$_{geo.}$, and (iii) pmf$_{bin.}$ bipartite degree distributions. The hyperbolic weighting is included here for comparison with Figure 6.

The binary and simple weighting schemes have no means of countering the impact of hub nodes when detecting communities, therefore they are often formed by the composition of multiple cliques. We also observe in Figure 6 that the binary and simple weighting schemes demonstrate the most variance across the 100 iterations suggesting a susceptibility to recording different results from different observations of the same process. Across each projection method we find uniformly high Gini coefficient among the distribution of detected community sizes (>0.75 for all p, weights and degree distributions). This shows that the range of community sizes is large, an unsurprising result given the agglomerative nature of the Louvain algorithm.

Figure 7 compares the hyperbolic, random walk, cosine, and Jaccard projections. We find that the four weighting schemes perform similarly in the case of geometric and binomial degree distributions for both the adjusted Rand index and mean community size. Modularity is similar among the random walk, cosine and Jaccard projections, but is still higher than that of the hyperbolic projection at low p. Considering the zeta degree distribution differentiates the projection schemes more clearly. Figure 7(b) shows that the cosine and Jaccard weightings perform poorly by reporting very high modularity when there is weak underlying structure in the network and showing little change as p increases. The random walk weighting reports some change in modularity as p increases. We also see that the hyperbolic weighting recovers the most information about the true network structure in Figure 7(a). In Figure 7(c), the Jaccard projection is unique in returning a structure with consistent expected community size across p and remains close to the value of the true partition. Figure 7 also shows that the cosine and Jaccard weighting schemes experience large variance over model observations, much like the binary and simple schemes. As with the first four projection methods, we find that the Gini coefficient for the random walk, cosine, and Jaccard weightings is uniformly high (>0.8 for all p, weights and degree distributions).

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

4. Discussion

Our exploration of the different community structures detected under the seven weighting schemes on the real-world dataset illustrates the huge impact that edge weighting has on community detection. As in most experimental cases, the true community structure is not known for the sharing of URLs on Twitter, so we cannot assess which is closest to some ground truth. Numerous previous studies strongly suggest that there is utility in this approach, so we are left in a position to decide which properties are desirable for further analysis, and assess the community quality by measuring coherence of some node properties.

The community structures reported on the Twitter dataset in Section 3.1 demonstrate the influence weighting schemes have on the resulting community partition. The Jaccard and binary weighting schemes stand out as performing poorly, resulting in one dominant community and a large number of small communities, certainly more than can be justified by a topical or demographic argument. If such a granularity is required for analysis, it is recommended that one of the other methods is applied alongside recursive community detection, that is subsequent use of community detection on the community subgraphs.

The cosine and random walk schemes perform differently from the other weightings, finding qualitatively different community structures. Neither finds any large communities, and the cosine weighting finds many more communities than other methods. This lack of similarity with the other methods suggests that the cosine and random walk weighting schemes encode different, less intuitive network properties than other projection schemes.

When applied to the Twitter dataset, the simple, hyperbolic, and unary weighting schemes perform similarly, finding similar communities both in size distribution and labeling. Under the unary weighting sizes initially decrease quickly, as a result it is likely that the simple or hyperbolic weighting schemes reflect an intuitive underlying community structure. Analysis of URL metadata (such as TF-IDF weighted importance of web domains within communities) supports this assertion by identifying qualitatively consistent communities formed around geographical or ideological factors.

Our exploration of synthetic networks covers a particular test case. We sample networks with approximately equal mode and community sizes and the same degree distribution for the left and right modes. Future work can expand on our analyses by permitting varying sizes and degree distributions in the modes. We exclude this work here given the combinatorially large search space for the various parameter combinations.

The adjusted Rand index scores in Figure 6 demonstrate that there is merit in using the unipartite projection process to identify community structure present in the bipartite network. Success with this method requires a sufficiently high community preference p to overcome the influence of the cliques formed by high-degree bipartite nodes on unipartite community detection. If the underlying level of community preference is too low, modularity maximizing community detection produces a poor representation of the bipartite network communities. This problem is exacerbated in degree distributions without a long tail; the existence of high-degree vertices within communities benefits the performance of community detection algorithms, and such vertices imply a long-tailed degree distribution.

It is important to note that compared to a baseline of zero, the modularity results found by community detection on the unipartite projections are deceptive as relatively high modularity is found even in the absence of *any* imposed bipartite community structure. In such circumstances, the projection process is creating local structures which are identified as spurious communities by modularity maximizing algorithms. Despite these concerns, there is evidence that unipartite community detection does recover information about the bipartite network; in the regime of high p, where community preference is strong, dense connections between cliques promote the identification of meaningful partitions through modularity optimization. Considering the standard null model with which network modularity is normally computed, the core assumption of edge

https://doi.org/10.1017/nws.2020.11 Published online by Cambridge University Press

independence is violated by the unipartite projection process; projection creates cliques rather than independent edges. It is possible that an adjusted null model which accounts for cliques may facilitate better community detection in projected networks.

Figure 7 shows that the random walk, cosine, and Jaccard projections perform similarly to the hyperbolic weighting in many cases but demonstrate some undesirable characteristics. The Jaccard and cosine weightings allow for partitions with very high modularity to be found even when there is no underlying community preference, particularly in the case of the zeta degree distribution. The random walk weighting does not suffer as much from modularity inflation, but requires a greater underlying preference to reproduce the true community structure with the same accuracy as the hyperbolic weighting.

Our experiments allow us to provide a recommendation for which of the seven projection methods studied is best overall. We frame such a recommendation in the experimental setting where the underlying community preference and structure are unknown and account for accuracy to the underlying structure, modularity of the optimal partition, and distribution of community sizes; these factors are all considered with their variance across the ensemble of model realizations. Our results demonstrate that the hyperbolic weighting scheme is the overall best method of the seven studied here, particularly for networks with long-tailed zeta degree distributions (as commonly found in socio-technical systems). As shown in Figure 6(a), the adjusted Rand index reveals that the hyperbolic weighting scheme most accurately recovers the bipartite community structure in nearly all cases, after a threshold of sufficiently strong community preference is passed. Hyperbolic weighting has the additional benefit of suppressing the inflated modularity scores common to many of the other methods and finding meaningful community sizes when a bipartite community structure exists to be found. The hyperbolic scheme also maintains small variance across the ensemble runs, suggesting more robustness to noise in the network. Beyond this optimal method, we also find that the binary and simple weighting schemes give qualitatively and quantitatively similar results in all experimental settings, suggesting that the simple weighting performs no better than an unweighted network.

Overall, this study of how community detection is affected by the edge weighting applied to the unipartite projection of bipartite networks with variable imposed community structure shows that careful thought needs to be given to the application of this approach and the interpretation of results. In terms of accuracy to the bipartite community structure, a useful direction would be to improve algorithms for community detection on bipartite networks directly; we note some recent efforts in this area (Zhou *et al.*, 2018). However, the projection approach is suitable in many circumstances. If a network arises through a growth process without preferential attachment, modularity maximizing community detection should be used carefully, as in these cases the detected modularity can be very high regardless of the underlying community structure. When the network growth process is driven by preferential attachment (as is the case in many real-world systems, and social networks in particular), the use of the hyperbolic weighting proposed by Newman (2001) generally finds the most accurate results to the true community structure. Modularity found on the unipartite projection cannot be thought of as directly representative of the community structure of the bipartite network as there are several weaknesses in the unipartite null model in this context. Future research into an alternative null model that better reflects the properties of projected networks (e.g. Arthur (2019)) would be of benefit to the wider scientific community given the widespread application of this approach when studying bipartite networks. An alternative direction for future work could apply the methods outlined here to explore how other measures of network structure are affected by edge weighting during the unipartite projection process.

Acknowledgments. TC is funded by an EPSRC Research Studentship (grant number EP/M506527/1). HW and IW acknowledge funding from ESRC (grant number ES/N012283/1) and NERC (grant number NE/P017436/1). The authors would like to acknowledge the use of the University of Exeter High-Performance Computing (HPC) facility in carrying out this work.

This manuscript extends earlier work presented at the Complex Networks 2018 conference in Cambridge, UK. We thank conference attendees for their insightful questions and comments which helped improve the research and the anonymous reviewers for their useful feedback. A list of tweet IDs used in Section 3.1 is available upon request.

Conflict of interest. Tristan J.B. Cann, Iain S. Weaver, and Hywel T.P. Williams have nothing to disclose.

Notes

1 https://developer.twitter.com/en/docs.html

References

Alzahrani, T., & Horadam, K. J. (2014). Analysis of two crime-related networks derived from bipartite social networks. In *Proceedings of the 2014 IEEE/ACM international conference on advances in social networks analysis and mining*, ASONAM 2014 (pp. 890–897).

Arthur, R. (2019). *Modularity and Projection of Bipartite Networks. arXiv e-prints*, arXiv:1908.02520.

Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, *76*(Dec), 066102.

Beckett, S. J. (2016). Improved community detection in weighted bipartite networks. *Royal Society Open Science*, *3*(1).

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Bongiorno, C., London, A., Miccichè, S., & Mantegna, R. N. (2017). Core of communities in bipartite networks. *Physical Review E*, *96*(Aug), 022321.

Callaway, D. S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. J., & Strogatz, S. H. (2001). Are randomly grown graphs really random? *Physical Review E*, *64*(Sep), 041902.

Cann, T. J. B., Weaver, I. S., & Williams, H. T. P. (2019). Is it correct to project and detect? Assessing performance of community detection on unipartite projections of bipartite networks. In *Complex networks and their applications VII* (pp. 267–279). Springer International Publishing.

Chen, Y.-Z., Li, N., & He, D.-R. (2007). A study on some urban bus transport networks. *Physica A: Statistical Mechanics and its Applications*, *376*, 747–754.

Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A., & Quattrociocchi, W. (2017). Mapping social dynamics on Facebook: The Brexit debate. *Social Networks*, *50*, 6–16.

Everett, M. G., & Borgatti, S. P. (2013). The dual-projection approach for two-mode networks. *Social Networks*, *35*(2), 204–210. Special Issue on Advances in Two-mode Social Networks.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378.

Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2007). Module identification in bipartite and directed networks. *Physical Review E*, *76*(Sep), 036102.

Isah, H., Neagu, D., & Trundle, P. (2015). Bipartite network model for inferring hidden ties in crime data. In *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015*. ASONAM 2015 (pp. 994–1001).

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLOS ONE*, *9*(6), 1–12.

Larivière, V., Gingras, Y., & Archambault, É. (2009). The decline in the concentration of citations, 1900–2007. *Journal of the American Society for Information Science and Technology*, *60*(4), 858–862.

Larremore, D. B., Clauset, A., & Jacobs, A. Z. (2014). Efficiently inferring community structure in bipartite networks. *Physical Review E*, *90*(Jul), 012805.

Lee, S., Song, S.-i., Kahng, M., Lee, D., & Lee, S.-g. (2011). Random walk based entity ranking on graph for multidimensional recommendation. In *Proceedings of the fifth ACM conference on recommender systems* (pp. 93–100). ACM.

Li, Y., & You, C. (2013). What is the difference of research collaboration network under different projections: Topological measurement and analysis. *Physica A: Statistical Mechanics and its Applications*, *392*(15), 3248–3259.

Marquitti, F. M. D., Guimarães, P. R., Pires, M. M., & Bittencourt, L. F. (2014). MODULAR: software for the autonomous computation of modularity in large network sets. *Ecography*, *37*(3), 221–224.

Melamed, D. (2014). Community structures in bipartite networks: A dual-projection approach. *PLOS ONE*, *9*(5), 1–5.

Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*, 016132.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351.

Romano, S., Vinh, N. X., Bailey, J., & Verspoor, K. (2016). Adjusting for chance clustering comparison measures. *Journal of Machine Learning Research*, *17*(134), 1–32.

Saracco, F., Straka, M. J., Di Clemente, R., Gabrielli, A., Caldarelli, G., & Squartini, T. (2017). Inferring monopartite projections of bipartite networks: an entropy-based approach. *New Journal of Physics*, *19*(5), 053022.

Sasahara, K. (2016). Visualizing collective attention using association networks. *New Generation Computing*, *34*(4), 323–340.

Schmidt, A. L., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2017). Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences*, *114*(12), 3035–3039.

Srivastava, A., Soto, A. J., & Milios, E. (2013). Text clustering using one-mode projection of document-word bipartite graphs. In *Proceedings of the 28th annual ACM symposium on applied computing*, *SAC 2013* (pp. 927–932).

Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *Proceeding of the international AAAI conference on web and social media*.

Wang, Z., Hou, T., Song, D., Li, Z., & Kong, T. (2016). Detecting review spammer groups via bipartite graph projection. *The Computer Journal*, *59*(6), 861–874.

Weaver, I. S. (2015). Preferential attachment in randomly grown networks. *Physica A: Statistical Mechanics and its Applications*, *439*, 85–92.

Williams, M., Cioroianu, I., & Williams, H. (2016). Different news for different views: Political news-sharing communities on social media through the UK General Election in 2015. In *Proceedings of the workshop on news and public opinion (NECO) at international AAAI conference on web and social media*.

Wyse, J., Friel, N., & Latouche, P. (2017) Inferring structure in bipartite networks using the latent blockmodel and exact ICL. *Network Science*, *5*(1), 45–69.

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326.

Zhou, C., Feng, L., & Zhao, Q. (2018). A novel community detection method in bipartite networks. *Physica A: Statistical Mechanics and its Applications*, *492*, 1679–1693.

Cite this article: Cann T. J. B., Weaver I. S., and Williams H. T. P. (2020). Is it correct to project and detect? How weighting unipartite projections influences community detection. *Network Science* **8**, S145–S163. https://doi.org/10.1017/nws.2020.11