Multilayer Modularity Belief Propagation

William Weir and Benjamin Walker

April 25, 2018

1 Abstract

2 Introduction

One popular heuristic for community detection involves optimizing a quantity developed by Newman and Girvan known as modularity [1]. Modularity compares the observed edges internal to the groups of a partition to the number expected under a random configuration model of the network. The formula for modularity modularity on a weighted network given by adjacency matrix $A_{i,j}$ is given by [1]

$$Q(\gamma) = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \,, \tag{1}$$

Where $k_i = \sum_j A_{i,j}$ is the weighted degree (strength) of node i, $m = \frac{1}{2} \sum A_{i,j}$ is the total edge weight of the network, and γ is the resolution parameter introduced by Reichardt and Bornholdt [2] that sets the scale of the communities identified. Setting $\gamma = 1$ gives the original modularity formulation in [1]. In general, maximizing the function over the combinatorially large space of possible partitions is *NP-Hard* and most methods attempt to find a locally optimal solution.

2.1 Modularity Belief Propagation

One of the crucial problems with the optimization of modularity as a means of community detection is that partitions of high modularity often exist even in completely random networks. Recently, Zhang and Moore applied the tools of statistical physics to overcome several of the challenges associated with previous heuristics aiming to maximize modularity[3]. They approach the problem as a spin-glass system with the energy given by $\mathcal{E} = -mQ(\{t\})$ and a Gibbs distribution by :

$$P(\{t\}) \propto e^{\beta \mathcal{E}(\{t\})} \tag{2}$$

where $\{t\} = [t_1, \ldots, t_N]; t_i \in \{1, ..., q\}$ being an assignment of the nodes in a network into q communities, and β representing an inverse temperature parameter determining how the probability is spread throughout the different states. As $\beta \to \infty$, the low temperature regime, only the partitions with the globally maximum modularity have non-zero probabilities. They suggest that instead of attempting to maximize modularity directly, community assignments should be made on the basis of the marginals of such a distribution in Eq. 2. There are several tools to compute the marginals of such the distribution such as Markov Chain Monte Carlo sampling, Gibb's sampling, and the class of algorithms called *Belief Propagation* (also known as the cavity method), for which the derive the following update conditions:

$$\psi_t^{i \to k} \propto \exp\left[\frac{\beta d_i}{2m}\theta_t + \sum_{j \in \partial ik} \log 1 + \psi_t^{j \to i}(\exp^{\beta} - 1)\right] \tag{3}$$

Where $\theta_t = \sum_j d_j \psi^j$ and ∂i denotes the neighbors of node i in the graph. We refer to Zhang and Moore's algorithm as *modbp*.

Zhang and Moore posit that in the event that if there exists several global maximum that are widely separated in the space of partitions (*i.e.* uncorrelated), belief propagation will oscillate between these partitions and fail to converge. The demonstrate this by showing the absence of a non-trivial retrieval phase for random ER graphs without structure.

2.2 Selection of Number of Communities

One issue with many community detection algorithms is in the selection of the appropriate number of communities. In the context of modularity, adjusting γ , "resolution parameter", in Eq. 1 can reveal communities of different scale and size therefore overcoming the "resolution limit of detection" first raised by [4]. Zhang and Moore do not include a resolution parameter in deriving their *modbp* algorithm (thereby implicitly setting $\gamma = 1$), and suggest an alternative approach to selecting the appropriate number of communities. They show in several examples that the maximum modularity achieved in the retrieval phase of the algorithm peaks at the appropriate number of communities, with no increase in $\hat{Q}$, the retrieval modularity once the appropriate number of communities is selected. However, this requires running *modbp* for many possible value of q, number of community assignments.

There have been two other approaches to selecting the appropriate number of communities using *modbp* without having to run the algorithm at many values of q. Both approaches involve selecting a $q_{\max}$, the largest possible number of communities, and then using the marginal probabilities of assignments to evaluate the true number of communities. Lai *et al.* noted that in the event that q is too large, many of the marginal community assignments will be highly correlated, and highly correlated states (community assignments) can

be condensed into a single group [5]. Similarly in Ref [6], they condense the community assignments on the basis of the average distance between the marginals across all nodes in the network. In practice, we have that choosing the number of communities the number of communities this way all but obliterates the retrieval phase if q_{max} is chosen too high above the actually number. We have implemented the method in Ref [6], letting the number of communities float up to a pre-specified $q_{\max}$ (see Methods), and show that incorporation of a resolution parameter, γ restores the width of the retrieval phase and returns closer to the correct number of communities.

2.3 Multilayer Modularity Belief Propagation

An extension of modularity for the multilayer case was developed by Mucha *et al.* to incorporated the coupling between the layers [7]. Using an analysis of modularity from a Laplacian dynamics perspective, they developed a null model for multilayer networks and gave the following formula for multilayer modularity:

$$Q(\gamma, \omega) = \sum_{i,j} \left(A_{ij} - \gamma P_{ij} + \omega C_{ij}\right) \delta(c_i, c_j) \tag{4}$$

While there are belief propagation models for the multilayer context, [8] for example, to date there has not been an extension of *modbp* to the multilayer context.

In this paper, we have extended Zhang and Moore's *modbp* method in three important ways. We allow for the presence of weighted edges, which can greatly influence the communities detected (see references needed). We have incorporated a resolution parameter, γ into the algorithm and show that this can create a wider retrieval phase and achieve better performance in the case where the number of communities is not known *a priori*. Finally, we have extended *modbp* to the multilayer framework developed by Mucha *et al.* [7] and demonstrated the use of this tool on both synthetic and real world data. We refer to our tool as *multimodbp* which can be used on both Multilayer networks and single-layer networks. We have developed a python package, implementing our method in a fast, efficient manner and interfaces with other standard networks tools.

3 Results

3.1 Single Layer

We begin by examining how our modifications effect the ability of *modbp* to detect communities within synthetically generated data in the single layer case. For single layer networks, our method collapses down to Zhang and Moore's with two main differences (see also **6. Method Description**):

1. We have included a resolution parameter, γ that adjusted the relative balance of the terms in the update equation. Like other implementations of modularity this effectively controls the size of the partitions identified within the retrieval phase.

2. We have set an upper limit on the number of communities, q_{max} and have incorporated the approach Ref [6] to select an effective number of communities based on the overlap of the marginal (see **2.2. Selection of Number of Communities**)

3.1.1 Single Layer Stochastic Block Model

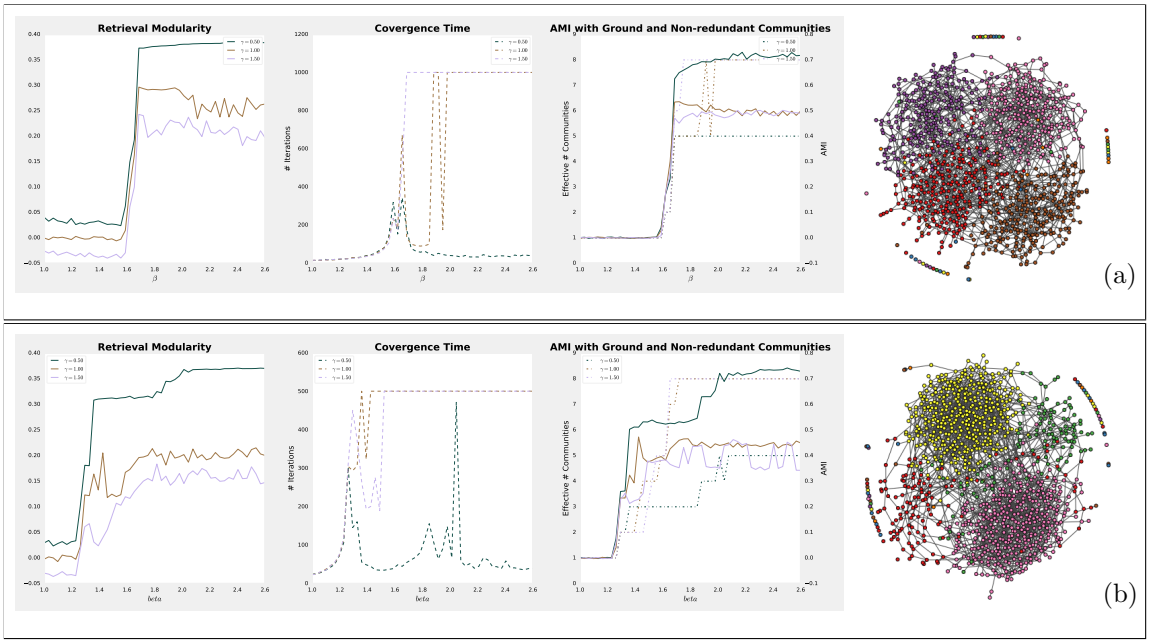

Figure 1: Demonstration of *multimodbp* on two realization of SBM model. From left to right the plots show the retrieval modularity, number of iterations to convergence, and the AMI of the retrieved partition with known community assignments and the effective number of communities. The black dotted lines denote the values of β^* for values of q ranging from 2 to $q_{\max}$ (see 6). **(a)** 4 community SBM with $n = 1000$, $\epsilon = \frac{p_{out}}{p_{in}} = .1$, $c_{avg} = 4$, and even community sizes and **(b)** and 4 community SBM with $n = 1000$, $\epsilon = .1$, $c_{avg} = 4$, with uneven community sizes ($\nu = [350, 150, 350, 150]$)

We examine the behavior of *multimodbp* on realization of a four community stochastic block model for different values of γ. First, we show that in the setting with several smaller communities, a higher value of γ produces a much wider retrieval phase and thus makes detection of communities more robust to selection of β. We generated a single realization of an SBM and scanned a range of β values to characterize the behavior of the algorithm seen in **Figure 1**. The retrieval modularity seen in **Figure 1a** shows a clear plateau for

$\gamma = .5$ (right most panel), corresponding with a broad retrieval phase (middle panel) that is absent at $\gamma = 1$. The AMI (left most panel) is clearly higher for all values of γ within the retrieval phase. What's more, the number of communities identified for $\gamma = .5$ plateaus at 5, which is more closely aligned with the underlying model (really there are only 4 main communities show in far right panel of **Figure 1a**).

We also tested the performance of the algorithm in the case where the sizes of the planted communities were uneven, shown in **Figure 1b**. The results here were similar though even more striking. There is a small retrieval phase for $\gamma = 1$, but it is much smaller than that of $\gamma = .5$ and the AMI is again consistently lower. For $\gamma = .5$ we actually detect two retrieval phases, the first one in which only nodes within the two larger communities are labelled correctly. Then as β increases the smaller two communities also become identifiable. This is consistent with the multiphase behavior observed in [6], though we note that in this example, the phase transition is only observed in a particular γ regime.

3.1.2 NCAA Division I-A College Football Network

We also look at how incorportation of the resolution parameter affects the ability of *multimodbp* to detection community structure on real world network. As

To look at how the value of γ effects the retrieval phase, we ran *multimodbp* for a range of γ value and examined the minimal number of iterations in the retrieval phase shown in **Figure 2**. For each value of γ, *multimodbp* was run over ten values in $\beta \in [1.5, 2.5]$, and values calculated at the β corresponding to the minimum number of iterations within this range. Runs that did not converge after 500 iterations suggest that for that value of γ the retrieval phase was either very small or nonexistent.

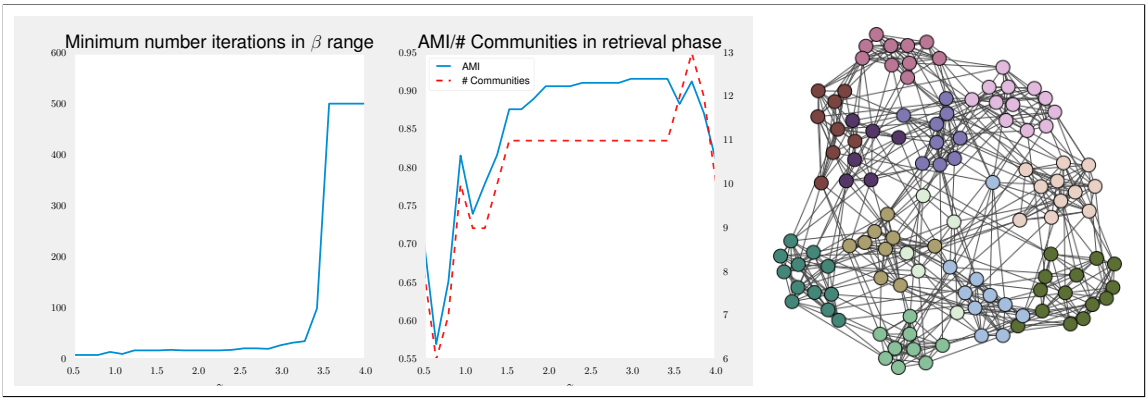

Figure 2

3.2 Multilayer

4 Discussion

5 Conclusion

6 Method Description

We have made two major modifications to Zhang and Moore's original modbp update equations. First, we have included a resolution parameter, γ which decouples the contribution of the two terms within the update equation. Second, we introduce an additional term to account for interlayer edges in a term that is nearly identical to the intralayer terms. It includes it's own interlayer coupling parameter, ω that influence's the relative contribution of this term.

$$\psi_t^{i\to k} \propto \begin{cases} \exp\left[\gamma\frac{\beta d_i}{2m}\theta_t + \sum_{j\in\partial i\setminus k}\log\left(1+\psi_t^{j\to i}(e^{\beta}-1)\right)\right] & (i,k)\in\mathcal{E}_{\text{intra}} \\ \exp\left[\sum_{j\in\partial i\setminus k}\log\left(1+\psi_t^{j\to i}(e^{\omega\beta}-1)\right)\right] & (i,k)\in\mathcal{E}_{\text{inter}} \end{cases} \tag{5}$$

Where $\mathcal{E}_{\text{inter}}$, denotes the interlayer edges, while $\mathcal{E}_{\text{intra}}$ is the set of interlayer edges. While, we have demonstrated *multimodbp* in the context of a specific multilayer topology, that is the multiplex network, our formulation is flexible enough to handle any type of multilayer networks with two classes of edges. In principle, the method could even be extended to many edge types, each with their own coupling parameter, ω_i.

6.1 Identification of β

$$\beta^* = \log\left(\frac{q}{(\sqrt{c}-1)}+1\right) \tag{6}$$

Where $c = \frac{<d^2>}{<d>} - 1$ is the excess degree. Based

References

[1] MEJ Newman and M Girvan. "Finding and evaluating community structure in networks". In: *Physical Review E* (2004).

[2] J Reichardt and S Bornholdt. "Statistical mechanics of community detection". In: *Physical Review E* (2006).

[3] Pan Zhang and Cristopher Moore. "Scalable detection of statistically significant communities and hierarchies, using message passing for modularity". In: *Proceedings of the National Academy of Sciences* 111.51 (2014), pp. 18144–18149.

[4] Santo Fortunato and Marc Barthelemy. “Resolution limit in community detection”. In: *Proceedings of the National Academy of Sciences* 104.1 (2007), pp. 36–41.

[5] Darong Lai, Xin Shu, and Christine Nardini. “Correlation enhanced modularity-based belief propagation method for community detection in networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 05.5 (May 2016), pp. 053301–.

[6] Christophe Schülke and Federico Ricci-Tersenghi. “Multiple phases in modularity-based community detection”. In: *Physical Review E* 92.4 (Oct. 2015), p. 042804.

[7] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. “Community structure in time-dependent, multiscale, and multiplex networks.” In: *Science* 328.5980 (May 2010), pp. 876–878.

[8] Amir Ghasemian, Pan Zhang, Aaron Clauset, Cristopher Moore, and Leto Peel. “Detectability Thresholds and Optimal Algorithms for Community Structure in Dynamic Networks”. In: *Physical Review X* 6.3 (July 2016), p. 031005.