Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

A Spin-Glass Model for Semi-Supervised Community Detection

Eric Eaton
Bryn Mawr College
Computer Science Department
Bryn Mawr, PA USA
eeaton@cs.brynmawr.edu

Rachael Mansbach
University of Illinois at Urbana-Champaign
Department of Physics
Urbana, IL USA
ramansbach@gmail.com

Abstract

Current modularity-based community detection methods show decreased performance as relational networks become increasingly noisy. These methods also yield a large number of diverse community structures as solutions, which is problematic for applications that impose constraints on the acceptable solutions or in cases where the user is focused on specific communities of interest. To address both of these problems, we develop a semi-supervised spin-glass model that enables current community detection methods to incorporate background knowledge in the forms of individual labels and pairwise constraints. Unlike current methods, our approach shows robust performance in the presence of noise in the relational network, and the ability to guide the discovery process toward specific community structures. We evaluate our algorithm on several benchmark networks and a new political sentiment network representing cooperative events between nations that was mined from news articles over six years.

1 Introduction

Many real networks, including social, financial, and biological networks, have natural community structures that are critical to functional and topological analysis. Automatic detection of these communities in relational networks has garnered interest in recent years with its success in a variety of applications. Current community detection methods (Newman & Girvan 2004; Newman 2006; Reichardt & Bornholdt 2006) automatically identify communities via analysis of the relational links between entities. The popular Newman-Girvan graph modularity (Newman 2006) is arguably the most widely used method for automatic community detection and the basis for many of these approaches. However, current modularity-based methods exhibit two key problems that complicate their application in many domains:

1. the inability to handle noise in the network, and
2. the tendency to admit a large number of high-scoring solutions without a clear optimum (Good et al. 2010).

The first problem results from the focus of current modularity-based methods to identify communities solely from analyzing the relationships between entities. In practical applications, the relational networks may be inaccurate or incomplete, confounding community detection using current algorithms. Since these methods rely on the accuracy of the relational network, their ability to discover the true community structures degrades rapidly as the network is perturbed by noise. In many cases, networks contain multiple overlapping communities, further complicating analysis.

The second problem is innate to using Newman-Girvan modularity to measure community partition quality. Exact community discovery is an NP-hard problem (Brandes et al. 2008), and so current algorithms return high-quality, rather than optimal solutions. The landscape of the Newman-Girvan modularity function typically admits a large (in some cases exponential) number of high-modularity solutions (Good et al. 2010). Consequently, although the modularity scores of these solutions may vary little, the communities themselves may be radically different. In many knowledge discovery applications, analysts may be interested in specific community structures or the application may impose constraints on the solution. For example, investigators tracing financial fraud may be interested in whether particular individuals are cycling cash between accounts to demonstrate money flow, or biologists may be interested in particular regulatory subgroups in a larger network. Current modularity-based algorithms cannot focus their search for community partitions that satisfy these requirements.

We address both of these problems by incorporating additional knowledge into the community detection process to both augment its performance in noisy networks and focus the discovery process on particular communities of interest. This paper develops a method for semi-supervised community detection, employing a spin-glass model from statistical physics to provide a rigorous foundation for combining external knowledge into the community detection process. The popular Newman-Girvan graph modularity reduces to a specific case of our model without the additional knowledge. We explore instantiations of our approach with two forms of external knowledge: labels on individual entities in the network, and pairwise constraints that specify the relevant community membership for pairs of entities. Effectively, the external knowledge focuses the community detection search on specific regions of the modularity landscape, both constraining and informing the solution.

Prior research has touched on the need to augment community detection with background knowledge, and has re-

sulted in several methods related to our approach. Our work is most similar to the investigation of semi-supervised community detection by Allahverdyan et al. (2010), which analyzed a spin-glass model where some of the spin-states are known and frozen in advance. In contrast to this work, our approach provides for deviation from the provided guidance if the support from the relational structure is strong enough, and we propose a formulation of community detection amenable to multiple forms of background knowledge. Our approach is also similar to the semi-supervised approach by Ma et al. (2010), which incorporates pairwise constraints into a symmetric nonnegative matrix factorization method for community detection, an alternative to modularity maximization. Also, the issue of focusing community detection on specific communities of interest has been briefly investigated by Hildrum and Yu (2005), who developed a method that grows the community model from a set of seed vertices to focus on specific regions of the network.

2 Automated Community Detection

We represent a relational network over a set of entities as an undirected weighted graph $G = (V, \mathbf{A})$ with vertices $V = \{v_1, v_2, \ldots, v_n\}$ and adjacency matrix $\mathbf{A}$, where $A_{ij} \in (0, 1]$ specifies that there is an edge e_{ij} between v_i and v_j with weight A_{ij}, and $A_{ij} = 0$ otherwise. The degree of vertex v_i is given by $d_i = \sum_j A_{ij}$, and the total weight of G is given by $m = \frac{1}{2}\sum_{i,j} A_{ij}$. Most current community detection methods seek to identify groups of vertices that are more densely connected within each community than between communities; see surveys by Fortunato (2010) and Namata et al. (2010) for comprehensive overviews.

The widely used Newman-Girvan *graph modularity* (Newman 2006) measures the community structure of the graph from a global perspective, gauging the differences of the graph's structure from an expected null model presumed to have no community structure. The modularity Q of a set of communities C in the network is given by

$$Q(C) = \frac{1}{2m}\sum\nolimits_{i,j}(A_{ij} - P_{ij})\delta(C_i, C_j) \ , \qquad (1)$$

where P_{ij} represents the probability of an edge between v_i and v_j in the null model, C_k represents the community to which v_k belongs, and $\delta(C_i, C_j)$ is the Kronecker delta function that is 1 if v_i and v_j belong to the same community (i.e., $C_i = C_j$) and 0 otherwise. Newman and Girvan employ a null model given by

$$P_{ij} = \frac{d_i d_j}{2m} \ , \qquad (2)$$

randomly rewiring the given graph while maintaining the total number of edges and the degree distribution of the vertices. High values of modularity Q indicate a strong community structure in the network, and Newman and Girvan describe several methods for identifying the communities using spectral clustering of the modularity matrix $(\mathbf{A} - \mathbf{P})$. Modularity has since been used as the foundation for a large number of other methods (Fortunato 2010).

Newman-Girvan modularity is a special case of another, more general measure of community structure based on the Potts spin-glass model from statistical mechanics (Reichardt & Bornholdt 2006). The Potts model is the multi-spin generalization of the classic two-state Ising spin-glass model, which is a collection of "up" and "down" spins in a graph configuration. Each spin state interacts only with the adjacent spin states in the graph, with an interaction energy that depends on whether the adjacent spins are alike or different. Each configuration of spins in the Ising model has a total energy associated with it, and minimizing the energy results in a stable ground state. For community detection, the ground state of the Ising model corresponds to splitting a graph into the two natural communities. The Potts spin model generalizes the Ising model to support q possible spin state values ($q \leq n$). If a graph contains t natural communities and $q \geq t$, then only t of the spin states will be populated in the ground state, enabling the Potts spin model to automatically identify the appropriate number of communities in the graph.

The ground state is found by minimizing the Potts model's Hamiltonian (Reichardt & Bornholdt 2006):

$$\begin{aligned}\mathcal{H}(C) = &-\sum\nolimits_{i\neq j} a_{ij}A_{ij}\delta(C_i, C_j) \\ &+ \sum\nolimits_{i\neq j} b_{ij}(1 - A_{ij})\delta(C_i, C_j) \\ &+ \sum\nolimits_{i\neq j} c_{ij}A_{ij}(1 - \delta(C_i, C_j)) \\ &- \sum\nolimits_{i\neq j} d_{ij}(1 - A_{ij})(1 - \delta(C_i, C_j)) \ .\end{aligned} \qquad (3)$$

The Hamiltonian rewards connections within communities and the lack of connections between communities, while penalizing for edges that disagree with this structure. The parameters a_{ij}, b_{ij}, c_{ij}, and d_{ij} balance the weights of these terms. We can view the spin-glass model as either examining in-degrees and out-degrees of nodes to identify communities (a local perspective), or examining deviations of the graph from a particular null model (a global perspective) similar to Newman-Girvan modularity. We can rewrite the Hamiltonian using a general probability function P_{ij} as

$$\mathcal{H}(C) = -\sum\nolimits_{i\neq j}(A_{ij} - \gamma P_{ij})\delta(C_i, C_j) \qquad (4)$$

by conveniently choosing $a_{ij} = c_{ij} = 1 - \gamma P_{ij}$ and $b_{ij} = d_{ij} = \gamma P_{ij}$ (Reichardt & Bornholdt 2006). This form of the Hamiltonian has an equivalent minimum to Eqn. 3 and closely matches the form of Newman-Girvan modularity given in Eqn. 1. In fact, Newman-Girvan modularity can be written as a specific case of the Potts spin-glass model by choosing P_{ij} via Eqn. 2 and normalizing the Hamiltonian.

3 Incorporating Guidance into Community Detection

While community detection has shown success when the relational network sufficiently captures the natural community structure, its success is limited in situations with missing entities and noisy relationships. In other cases, users may have specific communities of interest in mind or partial knowledge of the community memberships, requiring the search to focus on particular regions in the space of community partitions. In both these situations, additional knowledge can

inform the search, compensating for noise in the network or focusing discovery on particular community structures.

Although many forms of guidance are possible, we focus on methods that specify the community membership of individual entities or pairs of entities. We first develop a general method to incorporate these types of guidance into community detection, and then explore instantiations of this approach with specific forms of guidance in the next section.

Since the Hamiltonian (Eqn. 3) captures the total energy of the Potts model, we can incorporate external knowledge into community detection by penalizing for community structures that violate the guidance. In its general form, let the disagreement[1] of the communities to the given guidance be specified by a function $U : C \mapsto \mathbb{R}$. Although we focus on the general case for now, Section 4 discusses how to construct U for two specific types of guidance. We restrict U to be a function of the following form:

$$U(C) = \sum_{i \neq j} \Big(u_{ij} \big(1 - \delta(C_i, C_j)\big) + \overline{u}_{ij} \delta(C_i, C_j) \Big) \ , \quad (5)$$

where u_{ij} is the penalty for violating guidance that v_i and v_j belong to the same community (which increases $U(C)$ only when $C_i \neq C_j$), and $\overline{u}_{ij}$ is the penalty for violating guidance that v_i and v_j belong to different communities (which increases $U(C)$ only when $C_i = C_j$). Equivalently, Eqn. 5 can be written as

$$U(C) = \sum\nolimits_{i \neq j} \big(u_{ij} - (u_{ij} - \overline{u}_{ij}) \delta(C_i, C_j) \big) \ . \quad (6)$$

We can then incorporate guidance into the Hamiltonian as:

$$\mathcal{H}'(C) = \mathcal{H}(C) + \mu \sum_{i \neq j} \Big(u_{ij} - (u_{ij} - \overline{u}_{ij}) \delta(C_i, C_j) \Big) \ , \quad (7)$$

where $\mu \geq 0$ controls the balance between the inherent community structure and the external guidance. In this manner, U regularizes the Hamiltonian to control for deviation of the discovered communities from the provided guidance. The parameter μ could be set proportionally to the expected quality of the guidance, or to maximize performance either on a validation set or via cross-validation over the labeled data.

Following the assumptions Reichardt and Bornholdt (2006) used to derive Eqn. 4, the modified Hamiltonian (Eqn. 7) can be rewritten as

$$\begin{aligned} \mathcal{H}'(C) &= -\sum\nolimits_{i \neq j} (A_{ij} - \gamma P_{ij}) \delta(C_i, C_j) \quad (8) \\ &\quad + \mu \sum\nolimits_{i \neq j} \Big(u_{ij} - (u_{ij} - \overline{u}_{ij}) \delta(C_i, C_j) \Big) \\ &= -\sum_{i \neq j} \Big((A_{ij} - \gamma P_{ij}) \delta(C_i, C_j) \\ &\qquad + \mu (u_{ij} - \overline{u}_{ij}) \delta(C_i, C_j) \Big) + \mu \sum\nolimits_{i \neq j} u_{ij} \\ &= -\sum_{i \neq j} \Big(A_{ij} - \gamma \Big(P_{ij} - \frac{\mu}{\gamma} (u_{ij} - \overline{u}_{ij}) \Big) \Big) \delta(C_i, C_j) \\ &\quad + \mu \sum\nolimits_{i \neq j} u_{ij} \ , \quad (9) \end{aligned}$$

[1]The value of $U(C)$ is smaller when the communities C agree with the guidance, and larger when they disagree.

where P_{ij} is the probability of edge e_{ij} in the original null model. Note that the last term in Eqn. 9 is a constant for any partition, and so does not affect the optimization of $\mathcal{H}'(C)$. Therefore, we can drop it to reveal that Eqn. 9 is of the same form as Eqn. 4, measuring the cumulative deviation over all vertex pairs of A_{ij} to a modified null model $\mathbf{P}'$ given by

$$P'_{ij} = P_{ij} - \frac{\mu}{\gamma} (u_{ij} - \overline{u}_{ij}) \ . \quad (10)$$

Most importantly, we see that the guidance directly modifies the original null model $\mathbf{P}$ proportionally to the difference between the guidance to place v_i and v_j in the same community (i.e., u_{ij}) and the guidance to place them in different communities (i.e., $\overline{u}_{ij}$). In effect, the guidance reduces the null probability of edges between pairs of vertices that should be in the same community (when $u_{ij} > \overline{u}_{ij}$), and increases the null probability of edges for vertex pairs that should be in different communities (when $u_{ij} < \overline{u}_{ij}$).

In addition to being an intuitive way to guide community detection, this method provides a principled route to integrate external knowledge into the large number of modularity-based community detection methods by simply altering their null model. For example, choosing $\gamma = 1$ and $P_{ij} = \frac{d_i d_j}{2m}$ by Eqn. 2, then normalizing, we can derive the equivalent form of Newman-Girvan modularity that incorporates external guidance:

$$\begin{aligned} Q'(C) = \frac{1}{2m} \sum_{i \neq j} \left(A_{ij} - \left(\frac{d_i d_j}{2m} - \mu (u_{ij} - \overline{u}_{ij}) \right) \right) \delta(C_i, C_j) \\ - \frac{\mu}{2m} \sum\nolimits_{i \neq j} u_{ij} \ . \quad (11) \end{aligned}$$

The normalization constant $1/2m$ is included for convention following Newman (2006); it has no bearing on the optimization of the community structure since it is a constant, as is the last term of the expression. When the external guidance is either ignored (i.e., $\mu = 0$) or absent (i.e., $\forall_{i,j}\ u_{ij} = \overline{u}_{ij} = 0$), we recover Newman-Girvan graph modularity as a special case of our approach.

Although we analyze the Hamiltonian as Eqn. 9, we can rewrite it in a form more conducive to optimization. Starting with Eqn. 8, we can instead decompose it as

$$\begin{aligned} \mathcal{H}'(C) &= -\sum\nolimits_{i \neq j} (A_{ij} - \gamma P_{ij}) \delta(C_i, C_j) \\ &\quad + \mu \sum\nolimits_{i \neq j} \Big(u_{ij} - (u_{ij} - \overline{u}_{ij}) \delta(C_i, C_j) \Big) \\ &= -\sum_{i \neq j} M_{ij} \delta(C_i, C_j) - \mu \sum_{i \neq j} \Delta U_{ij} \delta(C_i, C_j) + K \quad (12) \end{aligned}$$

where $M_{ij} = A_{ij} - \gamma P_{ij}$, $\Delta U_{ij} = u_{ij} - \overline{u}_{ij}$, and all constant terms (which can be dropped from the optimization) are subsumed into $K = \mu \sum_{i \neq j} u_{ij}$. This form of the Hamiltonian allows us to compute the matrices $\mathbf{M}$ and $\mathbf{\Delta U}$ once, and then efficiently compute the objective function as the community assignments change during optimization. Since we need only compute these matrices once, our approach has the same computational complexity as standard modularity-based community detection.

4 Forms of Guidance

We focus on two forms of background knowledge that have shown success in the literature: individual labels and pairwise constraints. These forms provide meaningful guidance to the learning process and are intuitive for a user to specify.

4.1 Individual entity labels

The simplest form of guidance for community detection is labels placed on individual entities in the network. We assume that these labels are drawn from a hidden function $f : V \mapsto \mathbb{N}$ across G that respects the desired community structure. Entities which are given the same label are presumed to belong to the same community. The labeling over the graph is given by $Y = \{y_1, \ldots, y_n\}$, where $y_i \in \mathbb{N}_0$ (the set of natural numbers including 0) is the label for vertex v_i with $y_i = 0$ iff the label for v_i is unspecified.

We can then incorporate these labels into the Potts model by penalizing for entities with the same label being placed into different communities:

$$U(C) = \sum_{i \neq j} \mathbb{1}[y_i \neq 0 \wedge y_j \neq 0]\, \delta(y_i, y_j)\, (1 - \delta(C_i, C_j)) \ , \quad (13)$$

where $\mathbb{1}[p] = 1$ if predicate p is true and 0 otherwise, implying that

$$u_{ij} = \begin{cases} 1 & \text{when } y_i = y_j \neq 0 \\ 0 & \text{otherwise} \end{cases} \qquad \overline{u}_{ij} = 0 \ . \quad (14)$$

This formulation only penalizes for placing vertices with the same label in different communities. It does not, however, enforce that communities should be composed solely of vertices with identical labels. In penalizing only for separating vertices with the same label, we enable the optimization to combine multiple labels into one community, which can be useful for examining how different labeled groups interact with each other in the discovered community structure.

4.2 Pairwise constraints

From the constrained clustering literature (Wagstaff et al. 2001; Bilenko et al. 2004), pairwise constraints specify the relative cluster membership for pairs of entities. They can serve as an intuitive mechanism for users to identify pairs of entities that belong to either the same community (a *must-link constraint*) or different communities (a *cannot-link constraint*). A constraint $\langle v_i, v_j, w, type \rangle \in \mathcal{C}$ denotes the relative community membership for vertices v_i and v_j, with a non-negative cost $w \in \mathbb{R}_0^+$ of violating the constraint, and $type \in \{must\text{-}link, cannot\text{-}link\}$ indicating the type of constraint. For convenience, we denote the set of must-link and cannot-link constraints as $\mathcal{C}_{ml}$ and $\mathcal{C}_{cl}$ respectively. Users specify constraints by selecting two vertices, then selecting the type of constraint.[2] The value for w can be constant for all constraints, or optionally specified on a per-constraint basis proportionally to the user's confidence in the guidance. To improve the effectiveness of the given constraints, we also take the transitive closure of $\mathcal{C}_{ml}$ and $\mathcal{C}_{cl}$, and add the resulting constraints to the appropriate set.

[2]Note that using entity labels to define equivalence sets for deriving must-link constraints is equivalent to the formulation given in Section 4.1, provided that all constraints are weighted equally.

To incorporate constraints into community detection, we choose $U(C)$ to penalize for disagreements with the provided constraints $\mathcal{C}_{ml}$ and $\mathcal{C}_{cl}$. Following the metric labeling formulation of the generalized Potts model (Kleinberg & Tardos 2002) the total cost of disagreement is:

$$\begin{aligned} U(C) = \quad & \alpha_1 \sum_{\langle v_i, v_j, w_{ij} \rangle \in \mathcal{C}_{ml}} w_{ij} \big(1 - \delta(C_i, C_j)\big) \\ & + \ \alpha_2 \sum_{\langle v_i, v_j, \overline{w}_{ij} \rangle \in \mathcal{C}_{cl}} \overline{w}_{ij} \delta(C_i, C_j) \ , \end{aligned} \quad (15)$$

where α_1 and α_2 balance the contribution between must-link and cannot-link constraint violations. If we assume that $w_{ij} = \overline{w}_{ij} = 0$ for all pairs of vertices without a constraint, and assume that $w_{ij} > 0$ or $\overline{w}_{ij} > 0$ implies that the user has defined respectively a must-link or cannot-link constraint between v_i and v_j with the equivalent weight, we can rewrite Eqn. 15 to match the form of Eqn. 5, yielding:

$$\begin{aligned} U(C) &= \alpha_1 \sum_{i \neq j} w_{ij} \big(1 - \delta(C_i, C_j)\big) + \alpha_2 \sum_{i \neq j} \overline{w}_{ij} \delta(C_i, C_j) \\ &= \sum_{i \neq j} \Big(\alpha_1 w_{ij} \big(1 - \delta(C_i, C_j)\big) + \alpha_2 \overline{w}_{ij} \delta(C_i, C_j) \Big) \end{aligned}$$

and revealing that for pairwise constraints,

$$u_{ij} = \alpha_1 w_{ij} \qquad \overline{u}_{ij} = \alpha_2 \overline{w}_{ij} \ . \quad (16)$$

5 Evaluation

We evaluate our community detection approach in semi-supervised learning scenarios, considering its robustness under conditions of noise in the network and its ability to recover a specific community structure.

It has been widely shown that Newman-Girvan modularity performs very well for community detection when the relational data accurately reflects the community structures. However, many real-world networks, such as criminal networks and biological networks, contain incorrect or missing relationships due to the difficulty of obtaining complete data in these domains. Under such circumstances, the true community structures can be obscured in the networks, causing the performance of Newman-Girvan modularity to degrade. Our experiments show that incorporating background knowledge into community detection augments its performance in the presence of such noise in the network.

We also examine the ability of community detection to recover specific communities of interest. As shown by Good et al. (2010), Newman-Girvan modularity typically provides high scores for a large number of diverse community structures. Our results show that incorporating guidance into community detection focuses the search toward specific regions of the modularity landscape that both contain high-modularity solutions and agree with the provided guidance.

5.1 Methodology

In our evaluation[3], we follow Reichardt et al. (2006) and use simulated annealing (Kirkpatrick et al. 1983) to mini-

[3]An open source implementation of our algorithm is available on the first author's website, along with all relational networks and ground truth partitions used in the experiments.

mize the Hamiltonian (Eqn. 12). In the optimization, we use an initial temp of 10K with a multiplicative cooling factor of 0.985, returning the best state over 10 annealing trials. Candidate successor generation is done by stochastic community reassignment of a random entity. Also, for comparison with Newman-Girvan modularity and the large number of modularity-based methods, we set $\gamma = \alpha_1 = \alpha_2 = 1$, choose P_{ij} by Eqn. 2, and hold μ fixed at 1. For guidance provided in the form of pairwise constraints, we weight must-link and cannot-link constraints equally, setting $w_{ij} = \overline{w}_{ij} = 1$. Noise is added to the original relational networks by uniformly adding and deleting edges at random.

Our experiments analyze two benchmark networks for community detection under increasing levels of noise. The Doubtful Sound Dolphin network, as used by Lusseau et al. (2003), represents frequent associations between 62 dolphins living in Doubtful Sound, New Zealand. Following the temporary departure of a single key dolphin (named SN100), the dolphin community split into two smaller subgroups, which later reunited with the return of SN100. The Zachary Karate Club network (Zachary 1977) shows the friendships between the 34 members of a university karate club. The club later divided into two groups following an internal dispute among its members. In both of these benchmark networks, we know the resulting partitions, which we use as ground truth. The goal of community detection in both of these networks is to predict the ground truth communities given the relationships between entities.

We also examine a new Political Sentiment network that represents cooperative and hostile relationships between nations. It was created by identifying 336,555 distinct political events between 196 nations (or representatives of those nations) that were reported in news articles from January 1, 2005 through December 4, 2010. Each event had a corresponding Goldstein score (Goldstein 1992) assigned to it, which ranges from -10 (very hostile) to $+10$ (very cooperative) indicating the political character of the event. For each pair of nations that share events, we computed the average Goldstein score for cooperative ($\text{score} > 0$) acts, and eliminated nation pairs with less than six events total during that (approximately) six-year period. We then formed a network of 56 cooperative nations by connecting pairs of entities that had average Goldstein scores that were greater than or equal to 5.4, eliminating isolated vertices. This threshold of 5.4 was chosen from the Goldstein scale to identify cooperative relationships between nations that included substantial amounts of material support, instead of lesser cooperative actions, such as diplomatic agreements or policy support.

Since we do not have ground truth in the Political Sentiment network, we found the highest scoring community partition through extensive search, and measure against this partition for evaluation purposes. All constraints and labels are extracted from this community partition, allowing us to measure agreement to this particular community structure. The goal is to recover these specific communities from the network and background knowledge with less search and in the presence of noisy relationships.

Newman-Girvan modularity (Eqn. 1) provides a measure for the quality of the resulting communities with respect to the relational network without considering the ground truth partitions. For all experiments that add noise to the relational network, we measure the modularity of the resulting communities with respect to the *original* (non-noisy) graph in order to accurately assess the quality of the discovered communities independent of noise.

To measure the agreement of the discovered communities to either ground truth or a specific community structure, we use the pairwise F-measure (Basu 2005) – a variation of the information-theoretic F-measure adapted to measure the number of same-community pairs. The pairwise F-measure is the harmonic mean of precision and recall, given by

$$\textit{F-measure} = \frac{2 \cdot \textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}} \qquad (17)$$

$$\textit{precision} = \frac{|\mathcal{P}_{correct}|}{|\mathcal{P}_{pred}|} \qquad \textit{recall} = \frac{|\mathcal{P}_{correct}|}{|\mathcal{P}_{same}|} \ ,$$

where $\mathcal{P}_{pred}$ is the set of entity pairs predicted to be in the same community, $\mathcal{P}_{same}$ is the set of entity pairs actually in the same community, and $\mathcal{P}_{correct} = \mathcal{P}_{pred} \bigcap \mathcal{P}_{same}$ is the set of correct predictions.

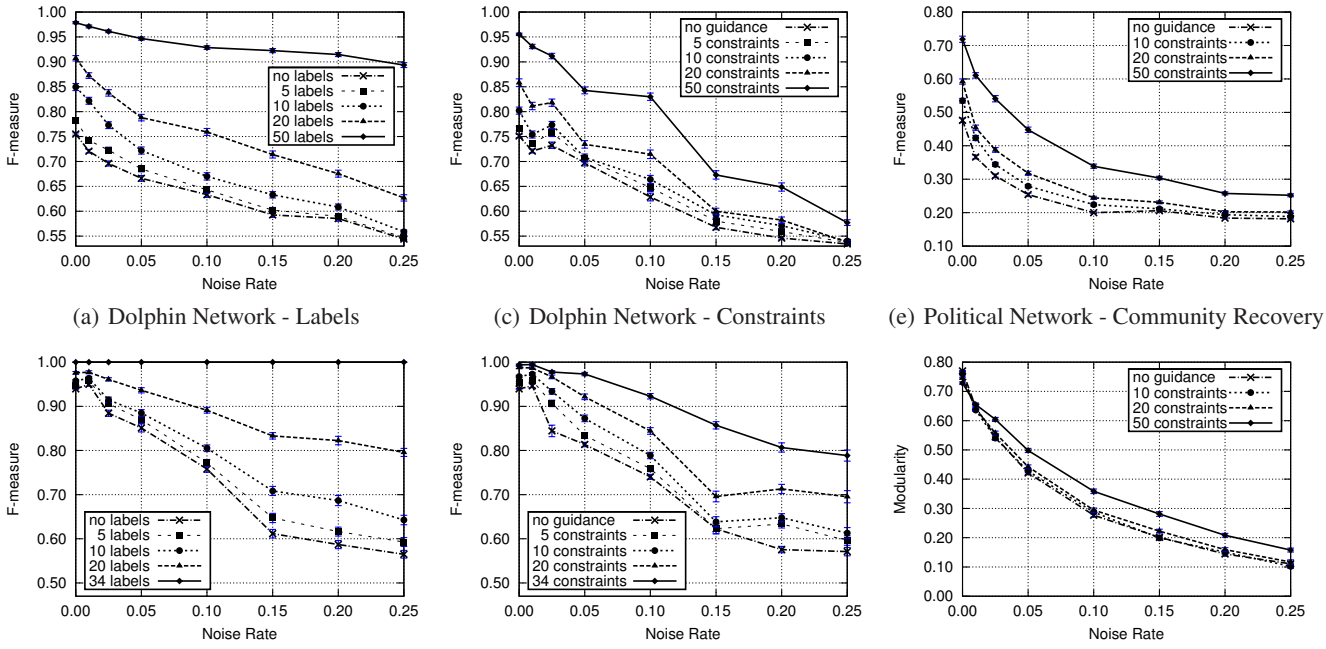
Our experiments examine the performance of community detection in semi-supervised learning, in which the background knowledge is given in batch to inform a single pass of community detection. We sample both the labels and pairwise constraints randomly from the ground truth communities. In the case of pairwise constraints, we extract must-link and cannot-link constraints in equal proportions.

5.2 Results

Figures 1(a)–(d) depict the performance of semi-supervised community detection with various amounts of guidance as the benchmark networks are perturbed by noise. With no noise, Newman-Girvan modularity[4] shows high performance. However, the addition of either labels or pairwise constraints only serves to improve performance over Newman-Girvan modularity. Adding even a small amount of guidance can significantly boost performance, even in the original (non-noisy) networks. For the dolphin network, the addition of only five constraints or labels increases the F-measure by 2–3%; the addition of ten constraints or labels improves the F-measure by 5–10%. For the karate network, we see a similar pattern, although to a lesser degree, with five constraints or labels increasing the F-measure by 1–2% and ten constraints or labels increasing it by 3–4%. Note that the karate network contains 34 vertices, and so providing one label for each vertex (Figure 1(b)) shows performance in the limit, which exactly recovers the ground truth community structure as we would expect.

As the amount of noise in the network increases, the background knowledge focuses the discovery process toward the target communities, compensating for the noisy relationships. Without guidance, the performance of community detection decreases severely as the amount of noise increases. The guidance serves to dampen the effect of noise, retaining higher performance. Therefore, community detection using

[4] In all figures, Newman-Girvan modularity is equivalent to our approach without guidance.



(a) Dolphin Network - Labels (c) Dolphin Network - Constraints (e) Political Network - Community Recovery

(b) Karate Network - Labels (d) Karate Network - Constraints (f) Political Network - Community Quality

Figure 1: The performance of semi-supervised community detection with various amounts of guidance as the relational networks are perturbed by noise. The noise rate refers to the proportion of random vertex pairs that have an edge either added or deleted. Figures (a)–(e) show the F-measure agreement to the target communities. Figure (f) shows the modularity of the discovered communities in the Political Sentiment network. The error bars, shown in blue, depict the standard error of the mean performances, which were averaged over 100 trials.

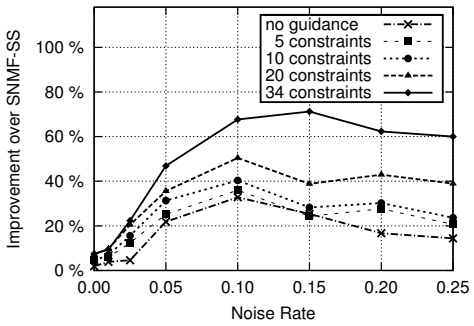
both the relational network and background knowledge may be most appropriate for applications where the communities may be obscured in noisy data, or where the network may contain erroneous connections.

Figures 1(e)–(f) examine the ability of community detection to recover a specific community structure in the Political Sentiment network using background knowledge. Recall that there are no ground truth communities for the Political Sentiment network, so we chose a specific partitioning with the highest modularity over an extensive search. Since modularity induces a number of high-modularity solutions, the task is then to recover the specific partitioning given background knowledge extracted from its community assignments. Additionally, the experiments did not employ as exhaustive of a search as the ground truth community discovery due to the practical time consideration of running the experiment over many trials and parameter settings, so the background knowledge also serves to focus the search on the highest region of the modularity landscape in less time.

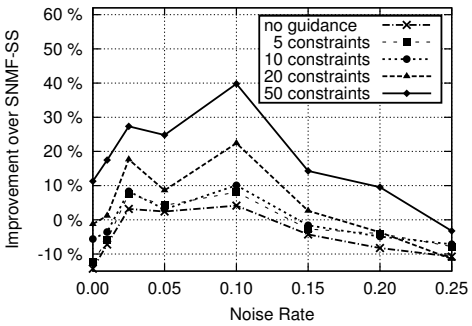
The results show that the background knowledge both focuses the community discovery process toward the target communities and, again, compensates for noise in the network. As Figure 1(f) reveals, all of the community partitions discovered without noise show high modularity, including the solutions which disagree with the target communities. However, Figure 1(e) shows that the addition of background knowledge yields solutions that increasingly match the target community structure. This ability to focus the solution returned by the community detection process is especially useful for applications with constraints on the acceptable solutions or for investigating specific communities of interest.

We also compare our approach to the SNMF-SS algorithm (Ma et al. 2010) for semi-supervised community detection in Figure 2. To make this a rigorous comparison, we tuned the parameters of SNMF-SS via line search to maximize performance on held-out data as measured over five trials of five-fold cross-validation on the noise-free networks. On the karate network, the performance of the two algorithms is roughly equivalent on the noise-free network. As the network becomes noisy, the performance of SNMF-SS rapidly degrades due to its reliance on the network structure. In comparison, our approach performs consistently better, achieving up to a 71% increase.

On the dolphin network, SNMF-SS initially performs better in noise-free scenarios with few constraints, but our approach shows significant improvement over SNMF-SS as the network becomes noisy or when additional constraints are supplied. In extreme noise scenarios, SNMF-SS again performs slightly better, but at this point, both algorithms are performing poorly due to the high noise (as shown in Figure 1(c)). The difference is further accentuated by SNMF-SS using parameters that were chosen in a noise-free setting, while our approach was tuned on the noisy network. Indeed, we found that when we required SNMF-SS to tune its parameters on the noisy network as well (as would be required in a real application with no knowledge of the true network), the performance difference between the algorithms became negligible in high-noise scenarios.



(a) Karate Network - Constraints



(b) Dolphin Network - Constraints

Figure 2: The percent improvement in F-measure of our approach against SNMF-SS (Ma et al. 2010) as the benchmark networks are perturbed by noise, averaged over 100 trials.

6 Conclusion

Incorporating background knowledge into the community detection process can significantly improve performance, especially in scenarios where the relational network contains noise. The background knowledge can also target the discovery process on specific communities of interest, compensating for the tendency of Newman-Girvan modularity to indiscriminately admit a large number of high-quality solutions. Both of these scenarios are likely to occur in deployed applications involving community detection, and our approach could serve as an important tool to improve community detection performance in such situations.

Acknowledgements

We would like to thank Jon Darvill and Paul Biancaniello of Lockheed Martin for providing the Political Sentiment data, and the anonymous reviewers for their feedback. This research was supported by ONR grant #N00014-10-C-0192.

References

Allahverdyan, A. E.; Steeg, G. V.; and Galstyan, A. 2010. Community detection with and without prior information. *EPL (Europhysics Letters)* 90(1):18002.

Basu, S.; Banerjee, A.; and Mooney, R. J. 2004. Active semi-supervision for pairwise constrained clustering. In *Proc. of the Int. Conf. on Data Mining*, 333–344. SIAM.

Basu, S. 2005. *Semi-Supervised Clustering: Probabilistic Models, Algorithms, and Experiments*. Ph.D. Dissertation, University of Texas at Austin.

Bilenko, M.; Basu, S.; and Mooney, R. 2004. Integrating constraints and metric learning in semi-supervised clustering. In *Proc. of the Int. Conf. on Machine Learning*, 81–88.

Brandes, U.; Delling, D.; Gaertler, M.; Görke, R.; Hoefer, M.; Nikoloski, Z.; and Wagner, D. 2008. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* 20(2):172–188.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486:75–174.

Goldstein, J. S. 1992. A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution* 36(2):369–385.

Good, B. H.; de Montjoye, Y. A.; and Clauset, A. 2010. Performance of modularity maximization in practical contexts. *Physical Review E* 81(4):046106.

Hildrum, K., and Yu, P. S. 2005. Focused community discovery. In *Proc. of the IEEE Int. Conf. on Data Mining*, 641–644. IEEE Press.

Kirkpatrick, S.; Gelatt, C. D.; and Vecchi, M. P. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Kleinberg, J., and Tardos, É. 2002. Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields. *Journal of the Association for Computing Machinery* 49(5):616–639.

Lusseau, D.; Schneider, K.; Boisseau, O. J.; Haase, P.; Slooten, E.; and Dawson, S. M. 2003. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology* 54(4):396–405.

Ma, X.; Gao, L.; Yong, X.; and Fu, L. 2010. Semi-supervised clustering algorithm for community structure detection in complex networks. *Physica A* 389:187–197.

Namata, G. M.; Sharara, H.; and Getoor, L. 2010. A survey of link mining tasks for analyzing noisy and incomplete networks. In *Link Mining: Models, Algorithms, and Applications*. Springer.

Newman, M. E. J., and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69(2):026113.

Newman, M. E. J. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* 74(3):036104.

Reichardt, J., and Bornholdt, S. 2006. Statistical mechanics of community detection. *Physical Review E* 74(1):016110.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schroedl, S. 2001. Constrained k-means clustering with background knowledge. In *Proc. of the Int. Conf. on Machine Learning*, 577–584. Morgan Kaufmann.

Zachary, W. W. 1977. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33:452–473.