Graph Transduction as a Non-Cooperative Game

Aykut Erdem

aykut.erdem@hacettepe.edu.tr

Faculty of Engineering, Hacettepe University, 06800 Beytepe, Ankara, Turkey

Marcello Pelillo

pelillo@dsi.unive.it

DAIS, Università Ca' Foscari Venezia, 30172 Venezia Mestre, Italy

Abstract

Graph transduction is a popular class of semi-supervised learning techniques which aims to estimate a classification function defined over a graph of labeled and unlabeled data points. The general idea is to propagate the provided label information to unlabeled nodes in a consistent way. In contrast to the traditional view, in which the process of label propagation is defined as a graph Laplacian regularization, this paper proposes a radically different perspective that is based on game-theoretic notions. Within the proposed framework, the transduction problem is formulated in terms of a non-cooperative multi-player game whereby equilibria correspond to consistent labelings of the data. An attractive feature of this formulation is that it is inherently a multi-class approach and imposes no constraint whatsoever on the structure of the pairwise similarity matrix, being able to naturally deal with asymmetric and negative similarities alike. Experiments

on a number of real-world problems demonstrate that the proposed approach performs well compared with state-of-the-art algorithms and it can deal effectively with various types of similarity relations.

1 Introduction

In the machine learning community, semi-supervised learning (SSL) has gained considerable popularity over the last decade (Chapelle et al., 2006; Zhu, 2005) and within the existing paradigms, graph-based approaches to SSL, namely the graph transduction methods, constitute an important class of algorithms. These methods model the geometry of the data as a graph with nodes corresponding to the labeled and unlabeled points and edges being weighted by the similarity between the points, and try to estimate the labels of unlabeled points by propagating the coarse information available at the labeled nodes to the unlabeled ones. Performing this propagation in a consistent way relies on a common a priori assumption, known as the *cluster assumption* (Zhou et al., 2004; Chapelle et al., 2006), which is reminiscent of the homophily principle used in social network analysis (Easley and Kleinberg, 2010). The assumption simply states that (1) points which are close to each other are expected to have the same label, and (2) points in the same cluster (or on the same manifold) are expected to have the same label. Building on this assumption, traditional graph-based approaches formalize graph transduction as a regularized function estimation problem on an undirected graph (Joachims, 2003; Zhu et al., 2003; Zhou et al., 2004).

This paper presents a novel game-theoretic perspective to the problem of graph transduction. Specifically, graph transduction is formulated in terms of a *multi-player non-cooperative game* where the players are the data points that take part in the game to decide their class memberships. In this setting, while the strategies played by the labeled points are already decided at the outset, as each of them knows which class it belongs to, the possible strategies available to unlabeled points are the whole set of hypotheses of being a member of one of the available classes. Within this formulation, the well-known Nash equilibrium concept for non-cooperative games provides a well-founded way of consistent labeling for the unlabeled points.

The game-theoretic interpretation presented in this paper is appealing for a number

of reasons. To begin with, it is intrinsically a multi-class approach. But more importantly, it can cope with both asymmetric and/or negative similarities. Since the seminal work of Tversky (1977), a large body of psychological studies suggests that human similarity judgments are non-metric (in particular, asymmetric). Non-metric similarities or dissimilarities also arise naturally in many practical applications, like comparing shapes (Jacobs et al., 2000) and protein sequences (Meila and Pentney, 2007). Common examples include the directed Hausdorff distance between sets, the Kullback-Leiber divergence between probability distributions, and Tversky's contrast model (Tversky, 1977; Santini and Jain, 1999). Non-metricity has been largely regarded as an artifact of poor choice of features or algorithms. Note, however, that it has been shown that non-metricity is sometimes essential to the nature of the problem and rendering the similarities metric may destroy relevant information (Laub et al., 2006).

The organization of the paper is as follows. Section 2 shows that transduction on an unweighted undirected graph can be formulated as a binary constraint satisfaction problem, which in turn provides the motivation for the proposed approach. Section 3 briefly reviews some basic notions of non-cooperative game theory. Next, Section 4 extends our analysis to the general case of transduction on weighted (directed) graphs, where graph transduction is formalized as a non-cooperative game using the connection between so-called relaxation labeling processes and game theory (Miller and Zucker, 1991). Section 5 investigates the loose relation between the proposed game-theoretic formulation and the traditional energy-based formulations. Section 6 reports experimental results on a number of real-world classification problems.

2   Transductive learning on unweighted undirected graphs

The theoretical motivation for the proposed approach stems from the analysis of the simplest case of graph transduction where the graph expressing the similarity relationships among the data points is an *unweighted undirected graph*. To give an example, the graph can be seen as a k-nearest neighbor (k-NN) graph with 0/1 weights over points in which the presence of an edge simply denotes the perfect similarity between a pair of two data points, otherwise the points are completely dissimilar. To illustrate this toy problem, consider the graph shown in Figure 1 in which edges are unweighted.

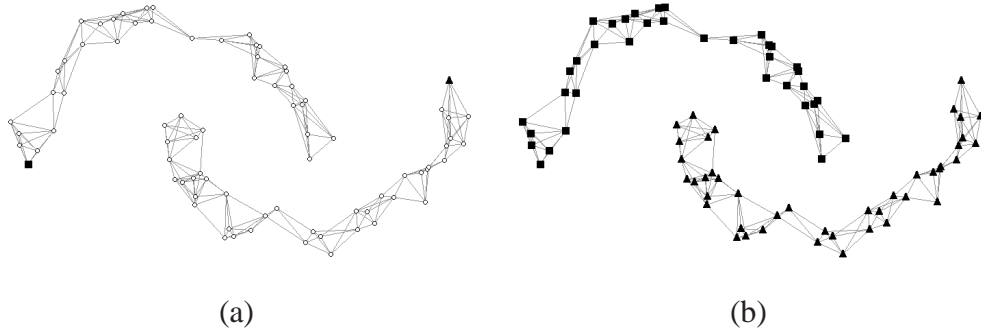<div align="center">(a)   (b)</div>

Figure 1: Transductive learning on an unweighted undirected graph. (a) An unweighted undirected graph describing the similarity relationships among the points. Two nodes of the graph are marked with different labels, respectively represented by a square and a triangle. (b) The consistent labeling of points satisfying the constraints based on the cluster assumption of SSL.

The classification task is to estimate the labels of the unlabeled points based exclusively on the information available at the two labeled points, each of which is marked with a different label. Recall the cluster assumption of semi-supervised learning that neighboring objects and objects in the same cluster (or on the same manifold structure) tend to belong to the same class. Clearly, in the unweighted graph setting, the cluster assumption is also valid and can be expressed as the hypothesis that every node in a connected component of a binary similarity graph has the same class label as each connected component describes a manifold.

Following this observation, we can formulate this toy version of graph transduction as a *(binary) constraint satisfaction problem* (CSP) (Tsang, 1993; Marriott and Stuckey, 1998). CSPs are widely used to solve combinatorial problems in a variety of application domains, such as artificial intelligence and computer vision. In the computer vision literature, the problem is often known as the *consistent labeling problem* (Waltz, 1975; Haralick and Shapiro, 1979).

A binary CSP is defined by a set of *variables* representing the elements of the problem being modeled and a set of binary *constraints* representing the relationships among variables. A solution of the problem is simply an assignment of values to the variables which satisfies all the constraints. If there is no such assignment, then the problem is unsatisfiable. When each variable can take a value from a finite *domain*, a *binary CSP* can be described in a formal manner as a triple (V, D, R), where $V = \{v_1, \ldots, v_n\}$ is a set of

variables, $D = \{D_{v_1}, \ldots, D_{v_n}\}$ is a set of domains of the variables, each D_{v_i} denoting a finite set of possible values for variable v_i, and $R = \{R_{ij} \mid R_{ij} \subseteq D_{v_i} \times D_{v_j}\}$ is a set of binary constraints, each R_{ij} describing compatible pairs of values for the variables v_i and v_j. If the cardinality of the domains of variables are p and q, respectively, then R_{ij} can be expressed by a 0/1 matrix of size $p \times q$, where $R_{ij}(\lambda, \lambda') = 1$ if the assignment $v_i = \lambda$ is compatible with the assignment $v_j = \lambda'$. For a general CSP on a finite domain, the problem of finding a solution is known to be NP-complete (Haralick et al., 1978). The simplest way to obtain an assignment satisfying all the given constraints or to report non-existence of such a solution is to perform backtracking. However, it is time consuming, so in practice, either a constraint propagation technique or a local search method is used to solve the problem (Tsang, 1993; Marriott and Stuckey, 1998).

Returning back to the motivating problem of transductive learning on an unweighted undirected graph, suppose that we are given a data set $\mathcal{D} = \{\mathcal{D}_\ell, \mathcal{D}_u\}$ consisting of labeled points $\mathcal{D}_\ell = \{d_1, \ldots, d_\ell\}$ and unlabeled points $\mathcal{D}_u = \{d_{\ell+1}, \ldots, d_n\}$ and a set of labels $\Phi = \{1, \ldots, c\}$ such that the labels provided for the first ℓ labeled points are given by $\{\phi_1, \ldots, \phi_\ell\} \in \Phi$. The task of transductive learning is to estimate the unknown labels $\{\phi_{\ell+1}, \ldots, \phi_n\}$ of unlabeled points $\{d_{\ell+1}, \ldots, d_n\}$. Now further suppose that the relationships among the data points are given by an unweighted undirected graph $\mathcal{G} = (\mathcal{D}, \mathcal{E})$, where $\mathcal{D}$ is the set of nodes and $\mathcal{E}$ is the set of edges such that an edge $e_{ij} \in \mathcal{E}$ shows that points i and j are perfectly similar to each other. Let $A = (a_{ij})$ denote the 0/1 adjacency matrix of $\mathcal{G}$. Reflecting the constraints imposed by the cluster assumption of SSL, the problem of graph transduction on an unweighted graph can be formalized as a binary CSP as follows:

- The set of variables: $V = \{v_1, \ldots, v_n\}$

- Domains: $D_{v_i} = \begin{cases} \{\phi_i\}, & \text{for all } 1 \leq i \leq \ell \\ \Phi & \text{for all } \ell + 1 \leq i \leq n \end{cases}$

- Binary constraints: $\forall ij$: if $a_{ij} = 1$, then $v_i = v_j$ $e.g.$ for a 2-class problem, $R_{ij} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Each assignment of values to the variables satisfying all the constraints is a solution of the CSP, and provides a consistent labeling for the unlabeled points.

Classical CSPs such as the one given in this section assume *crisp* constraints, in the sense that constraints are either completely satisfied or completely violated. However, for many real-world applications, such a formulation is too restrictive to be practical. A classical generalization to deal with soft constraints is described in (Hummel and Zucker, 1983), in which each constraint is assigned a weight representing a level of confidence. Later, it was shown that the notion of consistency proposed in (Hummel and Zucker, 1983) is related to the Nash equilibrium concept in non-cooperative game theory (Miller and Zucker, 1991). In this study, we build on this connection to devise a *graph transduction game* which serves as a generalization of the binary CSP for the motivating problem. For the sake of completeness, we next provide an overview of some basic notions from non-cooperative game theory (Nash, 1951).

3 Non-cooperative games and Nash equilibria

Non-cooperative game theory deals with models of strategic interactions (*games*) among anonymous agents (*players*), where the goal of each player is to maximize its own utility or payoff. Each player has a set of possible actions (*pure strategies*) to play, called the *pure strategy set*, and receives a payoff based on its own choice and those of the other players. In the definitions below, we restrict ourselves to *multi-player games* expressed in normal form and follow the notations used in (Weibull, 1995).

In normal form, a game with many players can be expressed as a triple $G = (\mathcal{I}, S, \pi)$, where $\mathcal{I} = \{1, \ldots, n\}$, with $n \geq 2$, is the set of *players*, $S = \times_{i \in \mathcal{I}} S_i$ is the *joint strategy space* defined as the Cartesian product of the individual pure strategy sets $S_i = \{1, \ldots, m_i\}$, and $\pi : S \to \mathbb{R}^n$ is the *combined payoff function* which assigns a real valued payoff $\pi_i(s) \in \mathbb{R}$ to each *pure strategy profile* $s \in S$ and player $i \in \mathcal{I}$. In the case of two players, payoff functions can be represented as two $m_1 \times m_2$ matrices. To illustrate, consider the well-known *rock-paper-scissors* game. It is a 2-player game in which each player has three possible strategies, *i.e.* $S_i = \{rock, \ paper, \ scissors\}$ for each $i = 1, 2$. Both players simultaneously choose a strategy and consequently receive a payoff based on their actions with respect to the rules "rock beats scissors", "scissors beats paper", and "paper beats rock". The payoff matrix of the game is given in Table 1 in which player 1 is the row player and player 2 is the column player. For example, if

6

player 1 plays *rock* and player 2 plays *paper*, player 1 loses the game, *i.e.* receives a payoff of -1, and player 2 wins the game, *i.e.* receives a payoff of 1.

	rock	*paper*	*scissors*
rock	(0,0)	(-1,1)	(1,-1)
paper	(1,-1)	(0,0)	(-1,1)
scissors	(-1,1)	(1,-1)	(0,0)

Table 1: The payoff matrix of the *rock-paper-scissors* game.

A *mixed strategy* of player $i \in \mathcal{I}$ is a probability distribution over its pure strategy set S_i, which can be described as the vector $x_i = (x_{i1}, \ldots, x_{im_i})^T$ such that each component x_{ih} denotes the probability that the player chooses to play its hth pure strategy among all the available strategies. Mixed strategies for each player $i \in \mathcal{I}$ are constrained to lie in the *standard simplex* of the m_i-dimensional Euclidean space $\mathbb{R}^{m_i}$:

$$\Delta_i = \left\{ x_i \in \mathbb{R}^{m_i} : \sum_{h=1}^{m_i} x_{ih} = 1, \text{ and } x_{ih} \geq 0 \text{ for all } h \right\}.$$

Accordingly, a *mixed strategy profile* $x = (x_1, \ldots, x_n)$ is defined as a vector of mixed strategies, each $x_i \in \Delta_i$ representing the mixed strategy assigned to player $i \in \mathcal{I}$, and each mixed strategy profile lives in the *mixed strategy space* of the game, given by the Cartesian product

$$\Theta = \times_{i \in \mathcal{I}} \Delta_i. \tag{1}$$

For the sake of simplicity, let $z = (x_i, y_{-i}) \in \Theta$ denote the strategy profile where player i plays strategy $x_i \in \Delta_i$ whereas the other players $j \in \mathcal{I} \setminus \{i\}$ play based on the strategy profile $y \in \Theta$, that is to say, $z_i = x_i$ and $z_j = y_j$ for all $j \neq i$. The expected value of the payoff that player i obtains can be determined by a weighted sum for any $i, j \in \mathcal{I}$ (when all the other players' strategies are kept fixed) as

$$u_i(x) = \sum_{s \in S} x(s) \pi_i(s) = \sum_{k=1}^{m_j} u_i \left(e_j^k, x_{-j} \right) x_{jk} \tag{2}$$

where $u_i \left(e_j^k, x_{-j} \right)$ denotes the payoff that player i receives when player j adopts its kth pure strategy, and $e_j^k \in \Delta_j$ stands for the *extreme mixed strategy* corresponding the vector of length m_j whose components are all zero except the kth one which is equal to

one. Note that for player j, playing its kth pure strategy is probabilistically equivalent to playing the extreme mixed strategy e_j^k.

The *mixed best replies* for player i against a mixed strategy $y \in \Theta$, denoted by $\beta_i(y)$, is the set of mixed strategies such that no other mixed strategy other than the ones included in this set gives a higher payoff to player i against strategy y:

$$\beta_i(y) = \{x_i \in \Delta_i : u_i(x_i, y_{-i}) \geq u_i(z_i, y_{-i}) \ \forall z_i \in \Delta_i\} \ .$$

Subsequently, the combined mixed best replies is defined as the Cartesian product of best replies of all the players $\beta(y) = \times_{i \in \mathcal{I}} \beta_i(y) \subset \Theta$.

Definition 1. *A mixed strategy $x^* = (x_1^*, \dots, x_n^*)$ is said to be a Nash equilibrium if it is the best reply to itself, $x^* \in \beta(x^*)$, that is*

$$u_i(x_i^*, x_{-i}^*) \geq u_i(x_i, x_{-i}^*) \tag{3}$$

for all $i \in \mathcal{I}, x_i \in \Delta_i$, and $x_i \neq x_i^*$. Furthermore, a Nash equilibrium x^* is called strict *if each x_i^* is the unique best reply to x^*, $\beta(x^*) = \{x^*\}$*

Nash equilibrium constitutes the key concept of game theory. It is motivated by the idea that a theory of rational decision-making should not be a self-destroying prophecy that creates an incentive to deviate for those who believe it. Indeed, the notion itself is a stability condition which states that no player can obtain a higher payoff by changing unilaterally its own strategy once such an equilibrium state is reached. Note that not all games do have a Nash equilibrium in pure strategies (e.g., the rock-paper-scissors game introduced above), but a fundamental result of game theory states that any normal-form game has at least one mixed Nash equilibrium (Nash, 1951). For example, the rock-paper-scissors game has a (unique) mixed Nash equilibrium which corresponds to the case where each player picks a strategy uniformly at random, *i.e.* $x_i = (1/3, 1/3, 1/3)$ for each $i = 1, 2$. The algorithmic issue of computing a Nash equilibria will be discussed later in Section 4.2.

4   The graph transduction game (GTG)

Consider the following *graph transduction game*. Assume each player $i \in \mathcal{I}$ participating in the game corresponds to a particular point in a data set $\mathcal{D} = \{d_1, \dots, d_n\}$

and can choose a strategy among the set of strategies $S_i = \{1, \ldots, c\}$, each expressing a certain hypothesis about its membership to a class and c being the total number of classes. Hence, the mixed strategy profile of each player $i \in \mathcal{I}$ lies in the c-dimensional simplex Δ_i. By problem definition, the players of the game can be categorized into two disjoint groups: those which already have knowledge of their membership, referred to as *labeled players* and denoted with the symbol $\mathcal{I}_\ell$, and those which do not have any idea about this at the beginning of the game, which are hence called *unlabeled players* and correspondingly denoted with $\mathcal{I}_u$.

The so-called labeled players of the game can further be distinguished based on the strategies they follow without hesitation, coming from their membership information. In formal terms, $\mathcal{I}_\ell = \{\mathcal{I}_{\ell|1}, \ldots, \mathcal{I}_{\ell|c}\}$, where each disjoint subset $\mathcal{I}_{\ell|k}$ stands for the set of players always playing their k^{th} pure strategies. It thus follows from this statement that each player $i \in \mathcal{I}_{\ell|k}$ plays its extreme mixed strategy $e_i^k \in \Delta_i$. In other words, x_i is constrained to belong to the minimal face of the simplex Δ_i spanned by $\{e_i^k\}$. In this regard, it can be argued that the labeled players do not play the game to maximize their payoffs since they have already chosen their strategies. In fact, the transduction game can be easily reduced to a game with only unlabeled players $\mathcal{I}_u$ where the definite strategies of labeled players $\mathcal{I}_\ell$ act as bias over the choices of unlabeled players.

It is important to note that any instance of the proposed transduction game will always have a Nash equilibrium in mixed strategies (Nash, 1951). Recall that, for the players, such an equilibrium corresponds to a steady state such that each player plays a strategy that could yield the highest payoff when the strategies of the remaining players are kept fixed, and it provides us a globally consistent labeling of the data set. Once an equilibrium is reached, the label of a data point (player) i is simply given by the strategy with the highest probability in the equilibrium mixed strategy of player i as

$$\phi_i = \arg \max_{h=1 \ldots c} x_{ih}, \tag{4}$$

thereby yielding a crisp classification.

4.1   Payoff functions

In this paper, we assume that the proposed graph transduction game is an instance of a special subclass of multi-player games, known as *polymatrix games* (Janovskaya,

1968; Howson, 1972), in which players are nodes of a graph and every edge denote a two-player game between corresponding pair of players. In other words, we suppose that only pairwise interactions are allowed in the game and the payoffs associated to each player are additively separable so that the payoff of each player is given by the sum of the payoffs gained from each game played with one of its neighbor. Formally speaking, for a pure strategy profile $s = (s_1, \ldots, s_n) \in S$, the payoff function of every player $i \in \mathcal{I}$ is in the form:

$$\pi_i(s) = \sum_{j=1}^{n} A_{ij}(s_i, s_j) \tag{5}$$

where A_{ij} is the *partial payoff* matrix between players i and j. It follows that, in terms of a mixed strategy profile $x = (x_1, \ldots, x_n)$, the payoffs are computed as $u_i(e_i^h) = \sum_{j=1}^{n} (A_{ij} x_j)_h$ and $u_i(x) = \sum_{j=1}^{n} x_i^T A_{ij} x_j$.

In an instance of the transduction game, since each labeled player is restricted to play a definite strategy of its own, all of these fixed choices can be reflected directly in the payoff function of a unlabeled player $i \in \mathcal{I}_u$ as follows:

$$u_i(e_i^h) = \sum_{j \in \mathcal{I}_{\mathcal{U}}} (A_{ij} x_j)_h + \sum_{k=1}^{c} \sum_{j \in \mathcal{I}_{\mathcal{D}|k}} A_{ij}(h, k) \tag{6}$$

$$u_i(x) = \sum_{j \in \mathcal{I}_{\mathcal{U}}} x_i^T A_{ij} x_j + \sum_{k=1}^{c} \sum_{j \in \mathcal{I}_{\mathcal{D}|k}} x_i^T (A_{ij})_k \tag{7}$$

Now, what is left is how to specify partial payoff matrices between each pair of players. For doing so, the binary logical constraints for the toy transduction problem can be simply replaced with their weighted versions. Let the geometry of the data be modeled with a weighted graph $\mathcal{G} = (\mathcal{D}, \mathcal{E}, w)$ in which $\mathcal{D}$ is the set of nodes representing both labeled and unlabeled points, and $w : \mathcal{E} \rightarrow \mathbb{R}$ is a weight function assigning a similarity value to each edge $e \in \mathcal{E}$. Representing the graph with its weighted adjacency matrix $W = (w_{ij})$, the partial payoff matrix between two players i and j is set as $A_{ij} = w_{ij} \times I_c$ where I_c is the identity matrix of size c. Note that when partial payoff matrices are represented in block form as $A = (A_{ij})$, the matrix A is given by the Kronecker product $A = I_c \otimes W$. Notice that in the case of binary-valued similarity relations, the above partial payoff matrices coincide with the compatibility matrices defined for the binary CSP given in Section 2. Further, if only pure strategies are allowed, then the transduction game reduces to the presented CSP. In a pure Nash equilibrium of such a game, the

neighboring players all play the same pure strategy to get the maximum support from their neighbors. Note that, however, when the data contains noise, then class manifolds may be connected; hence a pure Nash equilibrium may not exist, or in other words, the CSP may be unsatisfiable.

Empirically, we observed that specifying payoffs in terms of normalized similarity matrix $\widehat{W} = D^{-1/2}WD^{-1/2}$ with $D = (d_{ii})$ being the diagonal degree matrix of W whose elements are given by $d_{ii} = \sum_j w_{ij}$ performs better than the case with the original similarities. In that regard, we add that the use of normalization is a common practice in graph-based approaches because it can typically achieve a better performance. To give an example, while the GFHF method (Zhu et al., 2003) uses original (unnormalized) similarities, the LGC method (Zhou et al., 2004) employs the normalized similarity matrix in its formulation. Moreover, while not directly related to graph transduction, it has been shown that the use of normalization has nice convergence properties in spectral clustering (von Luxburg et al., 2004). In terms of game theory, however, it is interesting to note that both versions of the transduction game (with and without normalizing input similarities) belong to the so-called class of *normalized games*, *i.e.* games with payoffs in a unit-length interval (Daskalakis, 2011), but the gap in their classification performance requires further investigation.

4.2 Computing Nash equilibria

In the recent years, there has been a growing interest in the computational aspects of Nash equilibria. The general problem of computing a Nash equilibrium is shown to belong to the complexity class PPAD-complete, a newly defined subclass of NP (Daskalakis et al., 2009; Daskalakis, 2011). Nevertheless, there are many refinements and extensions of Nash equilibria which can be computed efficiently and moreover, the former result does not apply to certain classes of games (Nisan et al., 2007). This paper utilizes the well-established *evolutionary approach* (Weibull, 1995; Hofbauer and Sigmund, 1998), initiated by Maynard Smith (1982) to find a Nash equilibrium in a multi-player game, but we mention that other options do exist such as, for example, the simplicial subdivision method (van der Laan et al., 1987), continuation methods (Govindan and Wilson, 2003) and enumeration-of-support methods (Mangasarian, 1964; Dickhaut and

Kaplan , 1991; Porter et al., 2008).

The dynamic interpretation of Nash equilibria through the evolutionary approach imagines that the game is played repeatedly, generation after generation, during which a selection process acts on the multi-population of strategies, thereby resulting in the evolution of the fittest strategies. The selection dynamics is commonly modeled by the following set of ordinary differential equations:

$$\dot{x}_{ih} = g_{ih}(x)x_{ih} \tag{8}$$

where a dot signifies derivative with respect to time, and $g(x) = (g_1(x), \ldots, g_n(x))$ is the growth rate function with open domain containing $\Theta = \times_{i \in \mathcal{I}} \Delta_i$, each component $g_i(x)$ being a vector-valued growth rate function for player i. Hence, g_{ih} specifies the growth rate at which player i's pure strategy h replicates. It is generally required that the function g be *regular* (Weibull, 1995), *i.e.* (1) g is Lipschitz continuous and (2) $g_i(x) \cdot x_i = 0$ for all $x \in \Theta$ and players $i \in \mathcal{I}$. While the first condition guarantees that the system (8) has a unique solution through every initial state, the condition $g_i(x) \cdot x_i = 0$ ensures that the simplex Δ_i is invariant under (8).

The class of regular selection dynamics includes a wide subclass known as *payoff monotonic dynamics*, in which the ratio of strategies with a higher payoff increase at a higher rate. Formally, a regular selection dynamics (8) is said to be payoff monotonic if

$$u_i \left(e_i^h, x_{-i} \right) > u_i \left(e_i^k, x_{-i} \right) \Leftrightarrow g_{ih}(x) > g_{ik}(x) \tag{9}$$

for all $x \in \Theta$, $i \in \mathcal{I}$ and pure strategies $h, k \in S_i$.

A particular subclass of payoff monotonic dynamics, which is used to model the evolution of behavior by imitation processes, is given by

$$\dot{x}_{ih} = x_{ih} \left[\sum_{l \in S_i} x_{il} \left(\phi_i \left[u_i \left(e_i^h - e_i^l, x_{-i} \right) \right] - \phi_i \left[u_i \left(e_i^l - e_i^h, x_{-i} \right) \right] \right) \right] \tag{10}$$

where $\phi_i(u_i)$ is a strictly increasing function of u_i. The multi-population version of the replicator dynamics is obtained when ϕ_i is taken as the identity function, *i.e.* $\phi_i(u_i) = u_i$, as:

$$\dot{x}_{ih} = x_{ih} \left(u_i(e_i^h, x_{-i}) - u_i(x) \right) \tag{11}$$

The following theorem states that the fixed points of (11) are Nash equilibria.

Theorem 1. *A point $x \in \Theta$ is the limit of a trajectory of (11) starting from the interior of Θ if and only if x is a Nash equilibrium. Further, if point $x \in \Theta$ is a strict Nash equilibrium then it is asymptotically stable, additionally implying that the trajectories starting from all nearby states converge to x.*

Proof. See (Weibull, 1995). ☐

In the experiments, the discrete-time counterpart of (11) given below is utilized, where the mixed strategies of each unlabeled player is initialized to uniform probabilities, *i.e.* the barycenter of the simplex Δ_i:

$$x_{ih}(t+1) = x_{ih}(t)\frac{u_i(e_i^h)}{u_i(x(t))} \qquad (12)$$

The discrete-time replicator dynamics (12) has essentially the same dynamical properties as the continuous version (see, e.g., (Weibull, 1995) for a detailed analysis).

4.3 Computational Complexity

The computational complexity of finding a Nash equilibrium of a transduction game using (12) can be given by $\mathcal{O}(kcn^2)$, where n is the number of players (data points), c is the number of pure strategies (classes) and k is the number of iterations needed to converge. In theory, it is difficult to predict the number of required iterations, but experimentally, we noticed that it typically grows linearly on the number of data points[1]. Note that the complexity of popular graph transduction methods is also close to $\mathcal{O}(n^3)$. It is of interest that for two-player games, a fast evolutionary game dynamics has been proposed recently, which exhibits linear space and time complexity per iteration (Rota Bulò and Bomze, 2011). One can utilize this kind of dynamics to increase the efficiency of proposed approach.

5 Connection to graph-based approaches

Contrary to our derivation, the vast majority of graph-based SSL studies starts with writing down an objective function that casts the transductive learning problem as an

---

[1]In the case of asymmetric similarities, we have no Lyapunov function for the dynamics so convergence is not guaranteed. However, note that if the dynamics converges to a fixed point, it will definitely be a Nash equilibrium (Theorem 1).

energy minimization problem and most of the focus is on how to compute the optima of corresponding objective function. In general, all these methods attempt to estimate an optimal classification function which is defined on the nodes of the graph by minimizing an objective function with two terms. One term penalizes the mismatch between the initial label assignments and the labels estimated by the classifier. The second term is a regularization term that enforces the smoothness of the classification function. Although the game-theoretic perspective shifts the focus from optima of objective functions to equilibria of the non-cooperative games, in this section we try to shed some light on the connection between the proposed transduction game to the energy-based graph transduction methods. It is important to note that this analysis only investigates a special case in which the pairwise similarities are assumed to be symmetric and follows from the following property of polymatrix games.

Consider a polymatrix game with $A = (A_{ij})$ being the block matrix representation of partial payoff matrices between players, the average payoff for the whole population can be defined as:

$$E(x) = \sum_{i=1}^{n} x_i^T \left(\sum_{j=1}^{n} A_{ij} x_j \right) = x^T A x \tag{13}$$

The following proposition establishes a link between local maximizers of $x^T A x$ in Θ and Nash equilibria of polymatrix games with symmetric payoff matrices.

Proposition 1. *Suppose A is symmetric, that is $A_{ij} = A_{ji}$ for all $i, j \in \mathcal{I}$. Then, any local maximum $x^* \in \Theta$ of (13) is a Nash equilibrium point of the polymatrix game (Hummel and Zucker, 1983; Miller and Zucker, 1991).*[2]

Now consider a special instance of the proposed transduction game where the given pairwise similarities are symmetric, *i.e.* $w_{ij} = w_{ji}, \forall i, j \in \mathcal{D}$. Recall that graph transduction games belongs to the family of polymatrix games and its partial payoff matrices are all in the form $A_{ij} = w_{ij} \times I_c$. To compute Nash equilibria of a transduction game with symmetric similarities, one can use the above proposition and come up with the following constrained quadratic optimization problem by considering the special form

---

[2]Note, however, that this does not hold for the asymmetric case because $x^T A x$ is not a Lyapunov function for the dynamics.

of the partial payoff matrices:

$$\text{maximize } E(X) = \text{tr}\{X^T W X\}$$
$$\text{subject to } x_i \in \Delta_i \; \forall i \in \mathcal{I}_{\mathcal{U}} \tag{14}$$
$$x_i = e_i^k \; \forall i \in \mathcal{I}_{\mathcal{D}|k}$$

where $X = [x_1 \cdots x_n]^T$ is the $n \times c$ matrix of mixed strategies. Clearly, there is no guarantee that the solution found by the replicator dynamics (12) will be a global maximizer of (14). However, it was experimentally shown that, on a related problem, the basins of attraction of optimal or near-optimal solutions are quite large and the dynamics often converges to one of them (Pelillo, 1999; Pelillo et al., 1999). Moreover, it should be noted that even though globally optimal solutions are favorable as they yield the largest consistency, unlike standard approaches, in our game-theoretic interpretation, local optima have a meaningful interpretation and do indeed correspond to solutions of our problem, *i.e.* Nash equilibria of the transduction game.

For the special case above, we can now investigate its relation to energy-based formulations. To begin with, the functional in the above problem resembles the continuous relaxation of the *k-way normalized cut* criterion (Yu and Shi, 2003). However, note that there is a key difference in the game-theoretic formulation in that each mixed strategy x_i is constrained to lie in the c-dimensional standard simplex Δ_i. This subtle difference is very important since it provides robustness against noise and outliers (Pavan and Pelillo, 2007). Moreover, unlike the proposed approach, hard labeling constraints cannot be embedded into the Normalized Cuts framework in an explicit way, such that partial grouping constraints could be enforced by introducing extra linear equality constraints (Eriksson et al., 2007; Xu et al., 2009; Yu and Shi, 2004). The framework suggested recently in (Ghanem and Ahuja, 2010) is an exception but it is inherently a two-class clustering approach and requires a recursive strategy to solve multi-class problems.

In Section 4.1, we have defined partial payoff matrices for the graph transduction game as $A = I_c \otimes \widehat{W}$ where $\widehat{W} = D^{-1/2} W D^{-1/2}$ is the normalized similarity matrix. Suppose instead that they were specified as $A = I_c \otimes -L$, where $L = D - W$ is the unnormalized graph Laplacian. Then the resulting optimization problem becomes the

following:

$$\text{minimize } E(X) = \text{tr}\{X^T L X\}$$
$$\text{subject to } x_i \in \Delta_i \; \forall i \in \mathcal{I}_{\mathcal{U}} \tag{15}$$
$$x_i = e_i^k \; \forall i \in \mathcal{I}_{\mathcal{D}|k}$$

The above optimization problem is in fact equivalent to that of the graph Laplacian regularization used in (Zhu et al., 2003), which is known to be a special case of the regularization in (Zhou et al., 2004) with the parameter $\mu = \infty$ and the graph Laplacian being unnormalized. In this way, one can argue that the method in (Zhu et al., 2003) solves a special case of transduction games in which the pairwise similarities are symmetric and the partial payoffs are specified in terms of negative graph Laplacian.

6 Experimental Results

In this section, we analyze the performance and effectiveness of the proposed approach on some real-world classification problems. To highlight the property that the game-theoretic formulation can naturally deal with both symmetric, asymmetric and negative similarity relations alike, three groups of experiments are carried out. In each group of experiments, we compare our results against state-of-the-art graph-based semi-supervised learning algorithms and have obtained reasonably good results. It is important to note that none of the considered graph transduction methods can cope with all the three types of similarities.

6.1 Experiments with symmetric similarities

In this section, experiments are conducted on four real-world data sets: *USPS*[3], *YaleB* (Georghiades et al., 2001), *Scene* (Oliva and Torralba, 2001) and *20-news*[4]. Here are some details:

- *USPS* contains images of hand-written digits 0-9 down-sampled to 16×16 pixels and it has 7291 training and 2007 test examples. As in (Zhou et al., 2004), only

---

[3] http://www-stat.stanford.edu/~tibs/ElemStatLearn/
[4] http://people.csail.mit.edu/jrennie/20newsgroups/

the digits 1 to 4 from the training and test sets are selected, which gives a total of 3874 data points.

- *YaleB* is composed of face images of 10 subjects captured under varying poses and illumination conditions. As in (Breitenbach and Grudic, 2005), each image is down-sampled to 30×40 pixels and a subset of 1755 images are considered, which corresponds to the images of individuals 2, 5 and 8.

- *Scene* is a scene classification data set consisting of 2688 natural scene images classified into one of 8 classes. Each image is represented with a 512-dimensional GIST descriptor (Oliva and Torralba, 2001) which combines the outputs of Gabor-like filters specifically designed to capture the structural properties of a scene.

- *20-news* is the text classification data set used in (Zhou et al., 2004), which contains 3970 newsgroup articles selected from the 20-newsgroups data set, all belonging to the topic `rec` which is composed of the subjects `autos`, `motorcycles`, `sport.baseball` and `sport.hockey`. As described in (Zhou et al., 2004), each article is represented in 8014-dimensional space based on the TFIDF representation scheme.

Table 2 shows the summary of the data sets. For *USPS* and *YaleB*, each image pixel is treated as a single feature, thus each example is represented in 256-, and 1200-dimensional space, respectively. The similarity between two examples d_i and d_j is computed using the Gaussian kernel as $w_{ij} = exp(-\frac{dist(d_i,d_j)^2}{2\sigma^2})$ where $dist(d_i, d_j)$ is the distance between d_i and d_j and σ is the kernel width parameter. Among several choices for the distance measure $dist(\cdot)$, the Euclidean distance $\|d_i - d_j\|$ is evaluated for *USPS*, *YaleB* and *Scene*, and the cosine distance $dist(d_i, d_j) = 1 - \frac{\langle d_i,d_j \rangle}{\|d_i\|\|d_j\|}$ is evaluated for *20-news*.

	USPS	*YaleB*	*Scene*	*20-news*
# objects	3874	1755	2688	3970
# dimensions	256	1200	512	8014
# classes	4	3	8	4

Table 2: The data sets used in the experiments with symmetric similarities.

In the experiments, the proposed approach (denoted here with GTG) is compared against four well-known graph-based SSL algorithms, namely the Spectral Graph Transducer (SGT) (Joachims, 2003)[5], the Gaussian fields and harmonic functions based method (GFHF) (Zhu et al., 2003)[6], the local and global consistency method (LGC) (Zhou et al., 2004)[7] and Laplacian Regularized Least Squares (LapRLS) (Belkin et al., 2006)[8].

A crucial factor in the success of graph-based algorithms is the construction of the input graph as it represents the data manifold. As a result, graph construction for classification has become a subject of interest in recent years, *e.g.* (Jebara et al., 2009), but the problem is still open. To be fair in the evaluation, for all the methods, a fixed set of kernel widths is used and 9 different 20-NN graphs are generated by setting $w_{ij} = 0$ if x_j is not amongst the 20-nearest neighbors of x_i. In particular, the kernel width σ ranges over the set $linspace(0.1r, r, 5) \cup linspace(r, 10r, 5)$ with r being the average distance from each example to its 20th nearest neighbor and $linspace(a, b, n)$ denoting the set of n linearly spaced numbers between and including a and b.

Figure 2 shows the mean test error rates and one standard deviation error bars over 100 trials with different sizes of labeled data[9]. As it can be seen, LapRLS method gives the best results for the relatively small data sets, *YaleB* and *Scene*. However, for the other two, its performance is poor. In general, the proposed GTG algorithm is either the best or the second best algorithm. While its success is almost identical to that of the LGC method in *USPS*, *Yale-B* and *Scene*, it gives superior results for *20-news*.

6.2 Experiments with asymmetric similarities

In this section, several experiments are carried out on a real-world protein data set derived from *SCOP* (Structural Classification of Proteins) (Murzin et al., 1995)[10], and

---

[5]The optimal value of the parameter c is selected as the one with the best mean performance from the set $\{400, 800, 1600, 3200, 6400, 12800\}$.

[6]In obtaining the hard labels, the *class mass normalization* step is employed as suggested in (Zhu et al., 2003).

[7]As in (Zhou et al., 2004), the parameter α is set as 0.99.

[8]The optimal values of the extrinsic and intrinsic regularization parameters γ_A and γ_I is selected from the set $\{10^{-6}, 10^{-4}, 10^{-2}, 1\}$ for the best mean performance.

[9]We randomly select labeled samples such that each set contains at least one sample from each class.
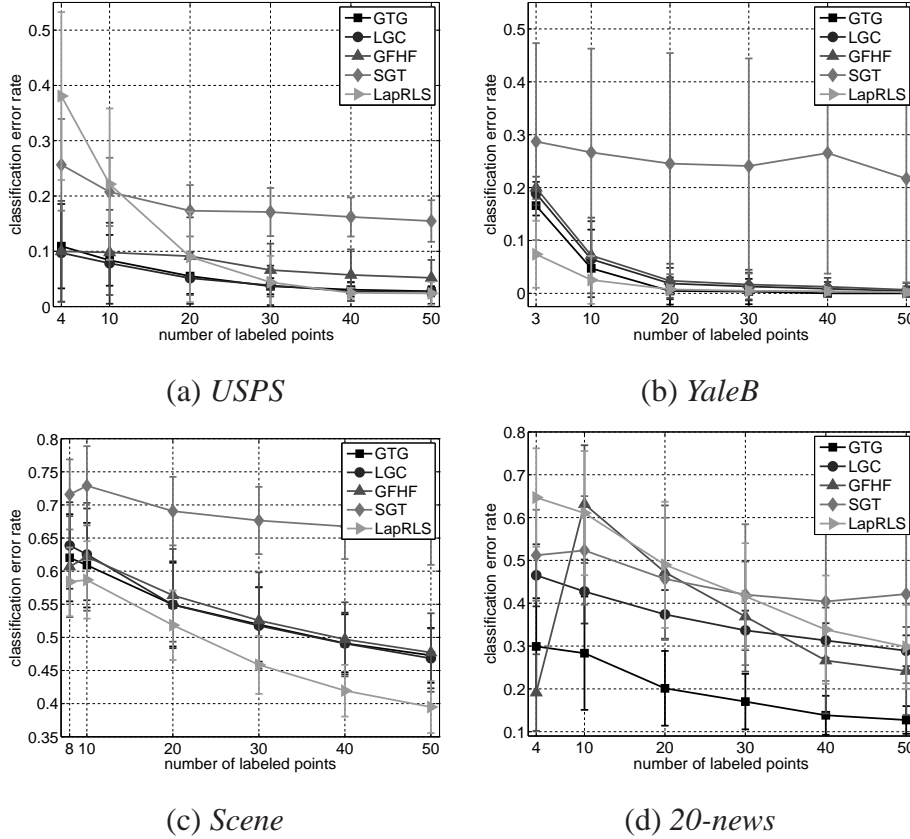
[10]Version 1.59 with less than 95% identity.

Figure 2: Performance comparisons on classification problems with *symmetric* similarities.

three document data sets – *Cora*, *Citeseer* (Sen et al., 2008)[11], and *WebKB*[12].

The details of these data sets are given below:

- *SCOP* contains 7329 proteins which are hierarchically divided into seven classes based on structural and evolutionary relationships. Each class is divided into folds, and each fold is further divided into superfamilies. Similar to the setup in (Meila and Pentney, 2007), only the proteins from the five largest folds of *all alpha* class are selected, which give a total of 451 protein sequences to be classified by fold.

- *Cora* contains 2708 machine learning publications classified into seven classes, and there are 5429 citations between the publications.

---

[11]Both the two data sets are available at `http://www.cs.umd.edu/projects/linqs/projects/lbc/`

[12]Available at `http://www.nec-labs.com/~zsh/files/link-fact-data.zip`

	SCOP	Cora	Citeseer	Cornell	Texas	Washington	Wisconsin
# objects	451	2708	3312	827	814	1166	1210
# classes	5	7	6	2	2	2	2

Table 3: The data sets used in the classification experiments with asymmetric similarities.

- *Citeseer* consists of 3312 scientific publications, each of which belongs to one of six classes, and there are a total of 4732 links.

- *WebKB* contains webpages collected from computer science departments of four universities (*Cornell*, *Texas*, *Washington* and *Wisconsin*), and each classified into seven categories. Following the setup in (Zhou et al., 2005), here we concentrate on classifying student pages from the others. Each subset respectively contains 827, 814, 1166 and 1210 webpages and 1626, 1480, 2218 and 3200 links.

The data sets are summarized in Table 3. For *SCOP*, E-values of the PSI-BLAST search calculated by (Weston et al., 2004)[13] are considered as the dissimilarity scores. These dissimilarity values are *not symmetric*. Instead of constructing NN graphs, this time the full similarity matrices are used where the kernel width σ was optimized for each method with respect to the set $linspace(0.5, 2.5, 5) \cup linspace(5, 20, 4) \cup 40$. For *Cora*, *Citeseer* and *WebKB*, as in (Zhou et al., 2005), only the citation/link structure is considered, even though one can also assign some weights by utilizing the textual content of the documents. Specifically, the experiments are performed on the link matrix $W = (w_{ij})$, where $w_{ij} = 1$ if document i cites document j and $w_{ij} = 0$ otherwise.

Unlike red the proposed game-theoretic approach, the standard methods mentioned before, namely SGT, GFHF, LGC and LapRLS, are subject to symmetric similarities. Hence, in this context, they can be applied only after rendering the similarities symmetric but this could result in loss of relevant information in some cases. In the evaluation, only the graph-based methods which can directly deal with asymmetric similarities are considered. Specifically, the proposed game-theoretic approach is compared against our implementation of the method in (Zhou et al., 2005), denoted here with LLUD.

---

[13]Available at`http://www.kyb.tuebingen.mpg.de/bs/people/weston/rankprot/`
`supplement.html`

This algorithm is based on the notion of random walks on directed graphs and it is equivalent to LGC in the case of symmetric similarities. It, however, assumes the input similarity graph to be strongly connected, so in (Zhou et al., 2005) the authors consider the *teleporting random walk (trw)* transition matrix as input, which is given by $P^\eta = \eta P + (1 - \eta)P^u$ where $P = D^{-1}W$ and P^u is the uniform transition matrix. This suggests a second variant for our framework, denoted with GTGtrw, where payoffs are defined in terms of this transition matrix. In the experiments, we fix $\eta = 0.99$ for both LLUD and GTGtrw. To provide a baseline, we also report the results of our approach that works on the symmetrized similarity matrices, denoted with GTGsym. For that case, we used the transformation $\widetilde{W} = 0.5 \times (W + W^T)$ for *SCOP*, and the symmetrized link matrix $\widetilde{W} = (\widetilde{w}_{ij})$ for the others, where $\widetilde{w}_{ij} = 1$ if either document i cites document j or vice versa, and $\widetilde{w}_{ij} = 0$ otherwise.

The test errors averaged over 100 trials are shown in Figure 3. Recall that the replicator dynamics used to find a Nash equilibrium is not guaranteed to converge in the case of asymmetric similarities. In fact, GTG succeeded to find a solution only on *SCOP* and failed on the others. However, we did not face any convergence problem with GTGtrw. Hence, we suspect that the this might be related to high sparseness of the data. Notice that the performances of GTGtrw and LLUD are quite similar on the classification problems in *WebKB* data sets. On the other hand, GTGtrw is superior in the multi-class problems in *SCOP*, *Cora* and *Citeseer*. It is important to note that symmetrization sometimes can provide good results. As shown in Figure 3(b)-(c), in *Cora* and *Citeseer*, GTGsym performs better than the other two methods.

6.3   Experiments with negative similarities

The game-theoretic formulation can also handle negative similarity or dissimilarity information in two-class semi-supervised classification problems in a natural way. In practice, such dissimilarity relationships could arise either in the computations, or they can be provided explicitly in terms of a set of cannot link pairs where the objects in each pair are expected to be assigned to different classes. The standard methods cannot accept negative similarities as well, since the existence of negative similarities could make their energy functions negatively unbounded. However, there are some graph-

(a) *SCOP*   (b) *Cora*   (c) *Citeseer*

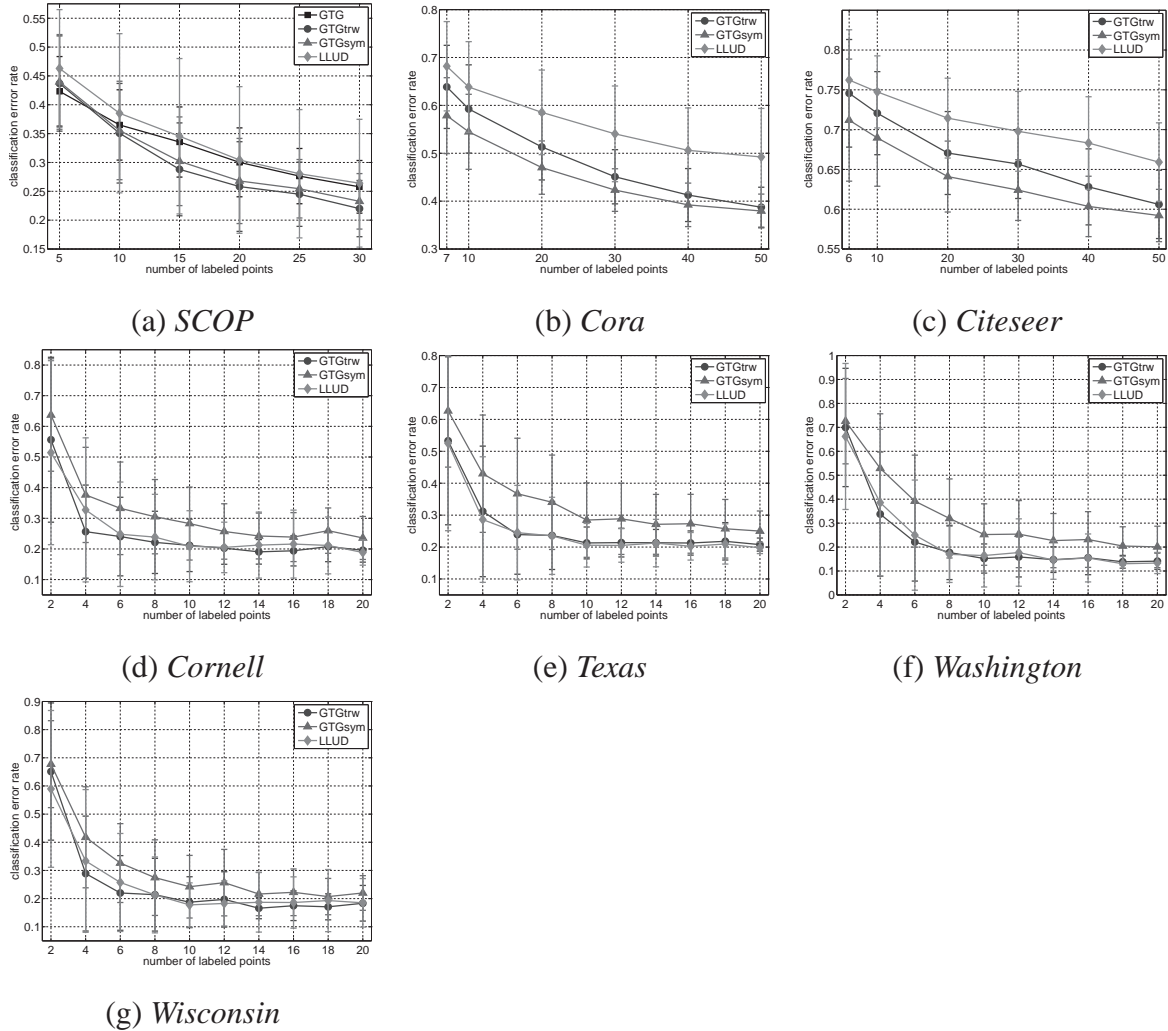(d) *Cornell*   (e) *Texas*   (f) *Washington*

(g) *Wisconsin*

Figure 3: Performance comparisons on classification problems with *asymmetric* similarities.

based SSL methods specifically designed for dealing with dissimilarity information, that is the mixed label propagation method (MLP) (Tong and Jin, 2007) and manifold regularization with dissimilarity method (MRWD) (Goldberg et al., 2007)[14], which is adopted from (Belkin et al., 2006). In this section, the proposed method is compared against these two algorithms.

The experiments are performed on 2 data sets from UCI repository[15]: *Ionosphere* and *Diabetes*. These data sets are described in Table 4. Notice that the similari-

---

[14]In the absence of negative similarities, MRWD reduces to LapRLS. In the experiments, we used the Regularized Least Squares (RLS) classifier and select the optimal values of the extrinsic and intrinsic regularization parameters as described for LapRLS.

[15]http://archive.ics.uci.edu/ml/

	Ionosphere	*Diabetes*
# objects	351	768
# dimensions	34	8
# classes	2	2

Table 4: The data sets used in the classification experiments with negative similarities.

ties derived from the data sets do not originally contain any negative values but we adopted the procedure in (Goldberg et al., 2007) and introduced *oracle* dissimilarity relations (cannot links) by randomly sampling pairs of examples having different labels. These pairs do not contain any labeled samples and to enforce the maximum degree of dissimilarity, the edge weights were set to the maximum similarity value exist in the data. In the experiments, the size of the labeled data is fixed is as 50 and the number of dissimilarity edges are varied between 3 and 12800. For all data sets, full similarity matrices are used and the kernel width σ is optimized with respect to the set $\{0.01\} \cup linspace(0.05, 0.25, 5) \cup linspace(0.25, 2.5, 10) \cup \{5, 10, 20, 25\}$.

The average test errors over 10 trials with randomly selected labeled examples and dissimilarity edges are given in Figure 4. As the methods in consideration explore both similarity and dissimilarity information, their accuracy improve as the size of the dissimilarity edges increases. While there is no considerable difference in the performances of the methods on *Diabetes*, GTG is clearly more successful on *Ionosphere*.
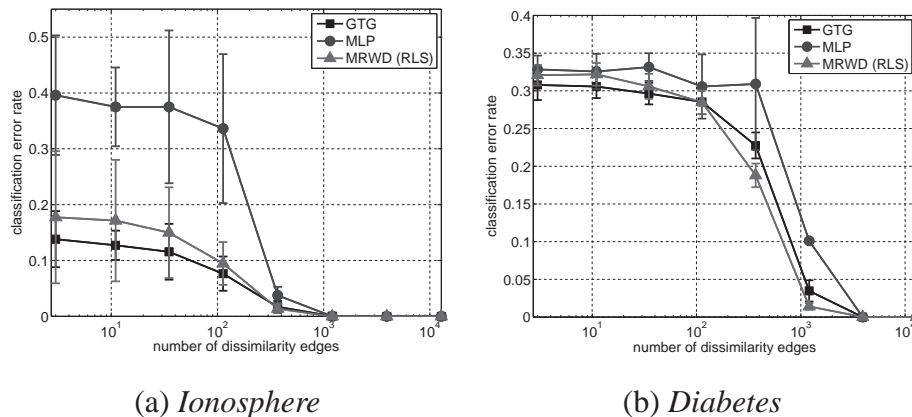

(a) *Ionosphere*   (b) *Diabetes*

Figure 4: Performance comparisons on classification problems with *negative* similarities

Conclusion

In this paper, a novel game-theoretic interpretation to graph transduction is introduced, in which the problem is formulated in terms of a polymatrix game whereby any equilibrium coincides with the notion of a consistent labeling of the data. As compared to existing approaches, the main advantage of the proposed framework is that there is no restriction on the pairwise relationships among data points; similarities and thus the payoffs can be negative or asymmetric. Apart from that, our approach is easy to implement and can be applied to multi-class problems. The experimental results show that the game-theoretic approach is not only more general but also competitive with standard approaches. As future work, we plan to focus on improving the efficiency. In our current implementation, we use the standard replicator dynamics to reach an equilibrium but we can study other selection dynamics that are much faster (Porter et al., 2008; Rota Bulò and Bomze, 2011). Another possible direction for future research is to generalize the presented approach to transductive learning in hypergraphs (Agarwal et al., 2006; Zhou et al., 2007). This will require replacing the pairwise interactions with higher-order interactions in defining payoffs, along the lines proposed in (Rota Bulò and Pelillo , 2009) for unsupervised learning. In this context, it would be especially interesting to explore whether different classes of games such as action-graph games (Jiang et al., 2011) will be more suitable or not.

Acknowledgments

References

Agarwal, S., Branson, K. & Belongie, S. (2006). Higher order learning with graphs. In *ICML*, 17–24.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, *7*, 2399–2434.

Breitenbach, M. & Grudic, G. Z. (2005). Clustering through ranking on manifolds. In *ICML*, 73–80.

Chapelle, O., Schölkopf, B., & Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Daskalakis, C. (2011). On the complexity of approximating a Nash equilibrium. In *SODA*.

Daskalakis, C., Goldberg, P. W., & Papadimitriou, C. H. (2009). The complexity of computing a Nash equilibrium. *Commun. ACM*, *52(2)*, 89–97.

Dickhaut, J., & Kaplan, T. (1991). A program for finding Nash equilibria. *Math. J.*, 87–93.

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds, and Markets : Reasoning About a Highly Connected World*. Cambridge University Press, Cambridge, UK.

Eriksson, A. P., Olsson, C., & Kahl, F. (2007). Normalized cuts revisited: A reformulation for segmentation with linear grouping constraints. In *ICCV*.

Georghiades, A. S., Belhumeur, P. N., & Kriegman, D. J. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, *23(6)*, 643–660.

Ghanem, B., & Ahuja, N. (2010). Dinkelbach NCUT: An efficient framework for solving normalized cuts problems with priors and convex constraints. *Int. J. Comput. Vis.*, *89(1)*, 40–55.

Goldberg, A., Zhu, X., & Wright, S. (2007). Dissimilarity in graph-based semi-supervised classification. In *AISTATS*.

Govindan, S., & Wilson, R. (2003). A global Newton method to compute Nash equilibria. *J. Econ. Theory*, *110*, 65-86.

Haralick, R.M., Davis, L.S., Rosenfeld, A., & Milgram, D.L. (1978). Reduction operations for constraint satisfaction. *Inf. Sci.*, 199–219.

Haralick, R. M., & Shapiro, L. G. (1979). The consistent labeling problem: Part I. *IEEE Trans. Pattern Anal. Mach. Intell.*, *1(2)*, 173–184.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, UK.

Howson, J. T. (1972). Equilibria of polymatrix games. *Management Science*, *18(5)*, 312–318.

Hummel, R. A., & Zucker, S. W. (1983). On the foundations of relaxation labeling processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, *5(3)*, 267–287.

Jacobs, D. W., Weinshall, D., & Gdalyahu, Y. (2000). Classification with nonmetric distances: Image retrieval and class representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, *22(6)*, 583–600.

Janovskaya, E. B. (1968). Equilibrium points in polymatrix games. *(in Russian) Litovskii Matematicheskii Sbornik*, *8*, 381–384 (Math. Reviews 39 #3831.).

Jebara, T., Wang, J., & Chang, S. (2009). Graph construction and *b*-matching for semi-supervised learning. In *ICML*, 441–448.

Jiang, A. X., Leyton-Brown, K., & Bhat, N. A. R. (2011). Action-graph games. *Games and Economic Behavior*, *71(1)*, 141–173.

Joachims, T. (2003). Transductive learning via spectral graph partitioning. In *ICML*, 290–297.

Laub, J., Roth, V., Buhmann, J. M. & Müller, K-R. (2006). On the information and representation of non-Euclidean pairwise data *Pattern Recognition*, *39(10)*, 1815–1826.

Mangasarian, O. (1964). Equilibrium points of bimatrix games. *J. Soc. Ind. Appl. Math.*, *12*, 778–780.

Marriott, K., & Stuckey, P. J. (1998). *Programming with Constraints: An Introduction*. MIT Press, Cambridge, MA.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK.

Meila, M., & Pentney, W. (2007). Clustering by weighted cuts in directed graphs. In *SIAM Conference on Data Mining (SDM)*.

Miller, D. A., & Zucker, S. W. (1991). Copositive-plus Lemke algorithm solves poly-matrix games. *Operations research letters*, *10(5)*, 285–290.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. `http://scop.berkeley.edu/`.

Nash, J. (1951). Non-cooperative games. *The Annals of Mathematics*, *54(2)*, 286–295.

Nisan, N., Roughgarden, T., Tardos, E., & Vazirani, V. V. (2007). *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.*, *42(3)*, 145–175.

Pavan, M., & Pelillo, M. (2007). Dominant sets and pairwise clustering. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, *29(1)*, 167–172.

Pelillo, M. (1997). The dynamics of nonlinear relaxation labeling processes. *J. Math. Imaging Vis.*, *7(4)*, 309–323.

Pelillo, M. (1999). Replicator equations, naximal cliques, and graph isomorphism. *Neural Computation*, *11(8)*, 1933–1955.

Pelillo, M., Siddiqi, K, & Zucker, S. W. (1999). Matching hierarchical structures using association graphs. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, *21(11)*, 1105–1120.

Porter, R. W., Nudelman, E., & Shoham, Y. (2008). Simple search methods for finding a Nash equilibrium. *Games and Economic Behavior*, *63(2)*, 642–662.

Rota Bulò, S., & Pelillo, M. (2009). A game-theoretic approach to hypergraph clustering. In *NIPS*, 1571–1579.

Rota Bulò, S., & Bomze, I. M. (2011). Infection and immunization: A new class of evolutionary game dynamics. *Games and Economic Behaviour (Special issue in honor of John F. Nash, jr.)*, *71*, 193–211.

Santini, S., & Jain, R. (1999). Similarity measures. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, *21(9)*, 871–883.

Sen, P., Namata, G. M., Bilgic, M., Getoor, L., Gallagher, B., & Eliassi-Rad, T. (2008). Collective classification in network data. *AI Magazine*, *29(3)*, 93–106.

Tong, W., & Jin, R. (2007). Semi-supervised learning by mixed label propagation. In *AAAI*, 651–656.

Tsang, E. (1993). *Foundations of Constraint Satisfaction*. Academic Press, London and San Diego.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

van der Laan, G., Talman, A., & van der Heyden, L. (1987). Simplicial variable dimension 1987 for solving the nonlinear complementarity problem on a product of unit simplices using a general labelling. *Math. Operations Res.*, *12(3)*, 377–397.

von Luxburg, U., Bousquet, O., & Belkin, M. (2004). On the convergence of spectral clustering on random samples: The normalized case. In *COLT*, 457–471.

Waltz, D. L. (1975). Understanding line drawings of scenes with shadows. In Winston, P. H., editor, *The Psychology of Computer Vision*, 19–92. McGraw-Hill.

Weibull, J. W. (1995). *Evolutionary Game Theory*. MIT Press, Cambridge, MA.

Weston, J., Elisseeff, A., Zhou, D., Leslie, C., & Noble, W. S. (2004). Protein ranking: From local to global structure in the protein similarity network. *Proc. of the National Academy of Science*, *101(17)*, 6559–6563.

Xu, L., Li, W., & Schuurmans, D. (2009). Fast normalized cut with linear constraints. In *CVPR*, 2866–2873.

Yu, S. X., & Shi, J. (2003). Multiclass spectral clustering. In *ICCV*, 313–319.

Yu, S. X., & Shi, J. (2004). Segmentation given partial grouping constraints. *IEEE Trans. Pattern Anal. Mach. Intell.*, *26(2)*, 173–183.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. In *NIPS*, 321–328.

Zhou, D., Huang, J., & Schölkopf, B. (2005). Learning from labeled and unlabeled data on a directed graph. In *ICML*, 1036–1043.

Zhou, D., Huang, J., & Schölkopf, B. (2007). Learning with hypergraphs: Clustering, classification, and embedding. In *NIPS*, 1601–1608.

Zhu, X. (2005). Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 912–919.