Graph Transduction as a Non-Cooperative Game

Aykut Erdem[1] and Marcello Pelillo[2]

[1] *Hacettepe University, Ankara, Turkey*
[2] *University of Venice, Venice, Italy*

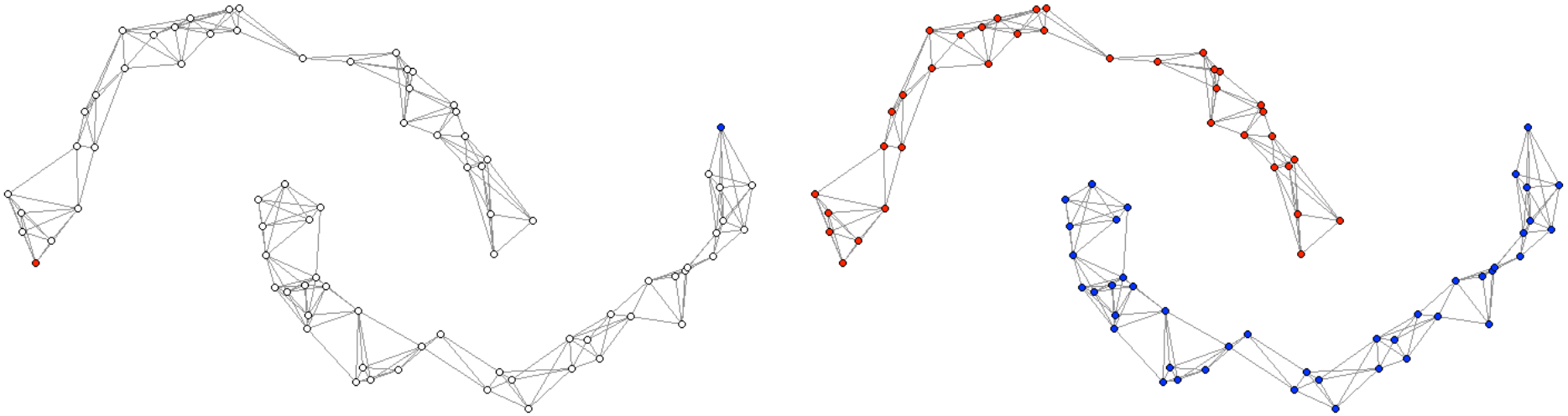
GBR 2011

Goal

- Development of a novel graph transduction method built upon a game-theoretic perspective
  - Graph transduction is formulated in terms of a non-cooperative multi-player game
  - Any equilibrium of the proposed game corresponds to a consistent labeling of the data.
- The proposed game-theoretic formulation imposes no constraint whatsoever on the structure of the pairwise similarity matrix
  - It naturally deals with symmetric, asymmetric and negative similarities alike.

Motivational problem: Transductive learning on unweighted undirected graphs

- The input graph G is an unweighted undirected graph
  - An edge denotes the perfect similarity between points
  - The adjacency matrix of G is a $0/1$ matrix



The cluster assumption: Each node in a connected component of the graph should have the same class label.

Motivational problem: Transductive learning on unweighted undirected graphs

- This toy problem can be formulated as a (binary) constraint satisfaction problem (CSP) as follows:
  - The set of variables: $V = \{v_1, \ldots, v_n\}$
  - Domains: $D_{v_i} = \begin{cases} \{y_i\} & \text{for all } 1 \le i \le l \\ Y & \text{for all } l+1 \le i \le n \end{cases}$
  - Binary constraints: $\forall i,j$: 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, providing a consistent labeling for the unlabeled points

Relaxation labeling, Non-cooperative games and Nash Equilibria

- Classical CSPs assume crisp constraints.
- *Relaxation labeling* is proposed to deal with soft constraints in which each constraint is assigned a weight representing a level of confidence.

  (Hummel and Zucker, 1983)
- The notion of *consistency* in relaxation labeling is in fact related to the *Nash equilibrium* concept in *non-cooperative game theory.*

  (Miller and Zucker, 1991)
- In this study, we used this connection to generalize the CSP formulated for the toy problem into a more general setting.

Basics of non-cooperative game theory

- Assume
  - a game between players *n* players $\mathcal{I}$
  - complete knowledge
  - a set of (pure) strategies $S_i = \{1, \ldots, m_i\}$ available to each player I
- Each player receives a payoff based on his own strategy and those of the other players.
- A mixed strategy of player i is a probability distribution over its strategies

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

Nash Equilibria

- Let $u_i(x_i, y_{-i})$ be the payoff obtained by player i playing x_i while the other players play according to profile $y_{-i} = (y_1, \ldots y_{i-1}, y_{i+1}, \ldots, y_n)$.

- **Definition.** A mixed strategy $x^* = (x_1^*, \ldots, x_n^*)$ is a Nash equilibrium if $u_i(x_i^*, x_{-i}^*) \geq u_i(x_i, x_{-i}^*)$ for all players i and $x_i \neq x_i^*$.

- Any non-cooperative game with finite set of strategies has at least one mixed Nash equilibrium (Nash, 1951).

Graph Transduction Game (GTG)

- Assume
  - the players participating the game correspond to the data points
  - the set of strategies available to each player denote the possible hypotheses about its class membership

  labeled players $\mathcal{I}_\ell = \{\mathcal{I}_{\ell|1}, \ldots, \mathcal{I}_{\ell|c}\}$,
  unlabeled players $\mathcal{I}_u$

- Labeled players choose their strategies at the outset.
  - each player $i \in \mathcal{I}_{l|k}$ always play its k^{th} pure strategy.

- The transduction game is, in fact, played among the unlabeled players to choose their memberships.
  - The label of an unlabeled player i is given by $y_i = \arg\max_{h \leq c} x_{ih}$

Defining payoff functions (1)

- Suppose that only pairwise interactions are allowed in the proposed game
  - an instance of a special subclass of multi-player games, referred to as polymatrix games (Janovskaya, 1968).

Polymatrix games

- Each player participates in a 2-player game with its neighbors
- The payoff of each player is given by the sum of partial payoffs from each game played with each of its neighbor

$$u_i(x) = \sum_{j=1}^{n} x_i^T A_{ij} x_j$$

Defining payoff functions (2)

- If the fixed choices of labeled players are considered, the payoff function is:

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

- But how to specify partial payoff matrices?
  - If $A = (A_{ij})$ represent partial payoff matrices in block form, we define $A = I_c \otimes W$

    e.g. for a 3 class problem, $A_{ij} = \begin{bmatrix} w_{ij} & 0 & 0 \\ 0 & w_{ij} & 0 \\ 0 & 0 & w_{ij} \end{bmatrix}$

We come up with a generalization of the binary CSP for the toy transduction problem!

Computing Nash equilibria

- To compute Nash equlibria, we used multi-population version of the replicator dynamics.

$$\dot{x}_{ih} = x_{ih}\left(u_i(e_i^h, x_{-i}) - u_i(x)\right) \qquad x_{ih}(t+1) = x_{ih}(t)\frac{u_i(e_i^h)}{u_i(x(t))}$$

continuous time (*) *discrete-time* (**)

Theorem. A point $x \in \Theta$ is the limit of a trajectory of (*) 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.

Connection to graph-based approaches

- Existing graph-based approaches cast transductive learning as an energy minimization problem.
  - The focus is on how to compute the optima of objective functions.
- The game-theoretic perspective shifts the focus from optima of objective functions to equilibria of games
  - There is no energy to minimize or maximize.
- We will analyze their connection for a special case in which the pairwise similarities are assumed to be symmetric.

A property of polymatrix games (with symmetric partial payoffs)

- 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:

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

Proposition. Suppose A is *symmetric*, that is $A_{ij}=A_{ji}$ for all players i,j. Then any local maximum $x^* \in \Theta$ of (*) is a Nash equilibrium point of the polymatrix game.

(Miller and Zucker, 1991)

Graph transduction game with symmetric similarities

- A Nash equilibrium of a transduction game with symmetric similarities ($w_{ij} = w_{ji}$ for all i,j) can be computed by solving the constrained quadratic optimization problem:

$$\begin{aligned} \text{maximize} \quad & E(X) = tr\{X^T W X\} \\ \text{subject to} \quad & x_i \in \Delta_i \quad \forall i \in \mathcal{I}_{\mathcal{U}} \\ & x_i = e_i^k \quad \forall i \in \mathcal{I}_{\mathcal{D}|k} \end{aligned} \qquad (**)$$

where $X = [x_1 \ldots x_n]^T$ is the matrix of mixed strategies.

- For this special subclass, we can now relate our approach with existing graph-based approaches.

Connection to Gaussian fields and harmonic functions method

- In the proposed approach, the partial payoff matrices are defined as $A = I_c \otimes W$.
- 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 (valid for only the symmetric case) becomes:

$$\begin{aligned} \text{minimize} \quad & E_{-L}(X) = tr\{X^T L X\} \\ \text{subject to} \quad & x_i \in \Delta_i \quad \forall i \in \mathcal{I}_{\mathcal{U}} \\ & x_i = e_i^k \quad \forall i \in \mathcal{I}_{\mathcal{D}|k} \end{aligned}$$

This problem is equivalent to that of the Gaussian fields and harmonic functions method (Zhu et al., 2003)

Experiments

- grouped into three based on the type of similarity relations:
  - symmetric similarities
  - asymmetric similarities
  - negative similarities

Symmetric similarities: Experimental Setting

- 4 data sets:
  - *USPS*, *YaleB*, *Scene*, *20-news*

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

- Methods compared:
  - *Gaussian fields and harmonic functions* (GFHF) (Zhu *et al.*, 2003)
  - *Spectral Graph Transducer* (SGT) (Joachims, 2003)
  - *Local and global consistency* (LGC) (Zhou *et al.*, 2004)
  - *Laplacian Regularized Least Squares* (LapRLS) (Belkin *et al.*, 2006)

- Gaussian kernel

$$w_{ij} = \exp\left(-\frac{dist(d_i, d_j)^2}{2\sigma^2}\right)$$

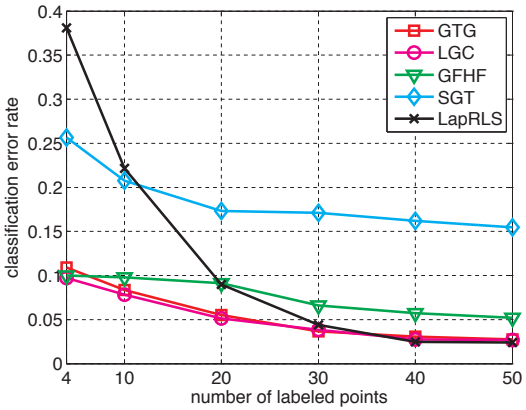
for *USPS*, *YaleB*, *Scene:* $dist(d_i, d_j) = \|d_i - d_j\|$

for *20-news:* $dist(d_i, d_j) = 1 - \frac{\langle d_i, d_j \rangle}{\|d_i\|\|d_j\|}$

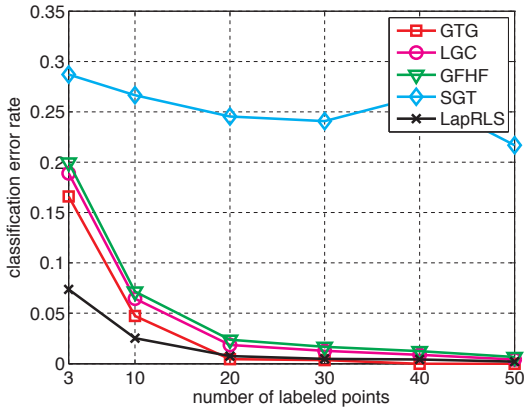
- 9 different 20-NN graphs

$\sigma \in linspace(0.1r, r, 5) \cup linspace(r, 10r, 5)),$ r: the average distance from each example to its 20th NN

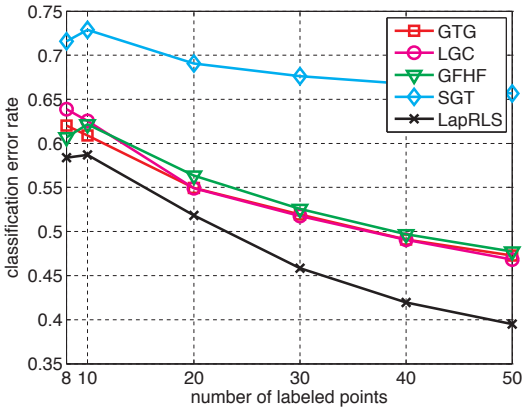
Symmetric similarities: Results



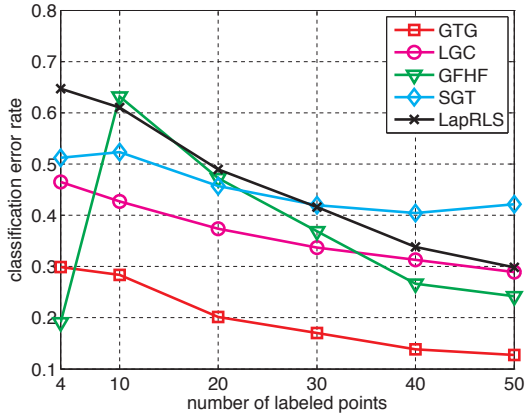
(a) *USPS*



(b) *YaleB*



(c) *Scene*



(d) *20-news*

The average test errors over 100 trials with different sizes of labeled data

Asymmetric similarities: Experimental Setting (1)

- Data sets:
  - *SCOP* (Structural Classification of Proteins)
  - *Cora*, *Citeseer*, *WebKB* (*Cornell*, *Texas*, *Washington*, *Wisconsin*)

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

- For *SCOP*
  - Dissimilarity scores are given by the E-values of the PSI-BLAST search (Weston *et al.*, 2004)
  - Gaussian kernel, 10 different candidate input graphs (full similarities)

$$\sigma \in linspace(0.5, 2.5, 5) \cup linspace(5, 20, 4) \cup \{40\}$$

- For *Cora*, *Citeseer* and *WebKB*
  - Only the citation/link structure is considered (Zhou *et al.*, 2005)

Asymmetric similarities: Experimental Setting (2)

- Classical approaches are subject to symmetric similarities!
- Method compared:
  - Baseline: GTG on symmetrized similarities (GTGsym)

$$\widetilde{W} = 0.5 \times (W + W^T) \quad \text{for } \textit{SCOP}$$
$$\widetilde{W} = min(W + W^T, 1) \quad \text{for others}$$

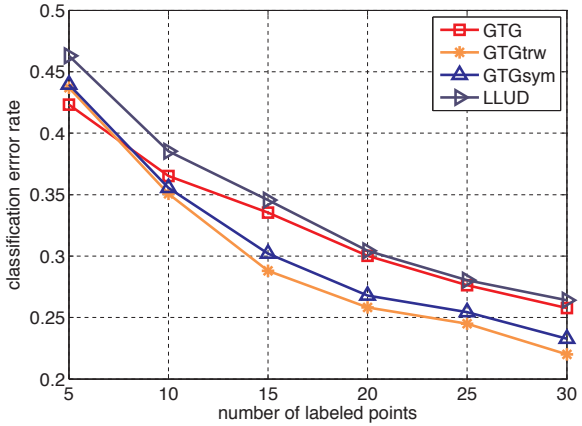
  - *Learning from labeled and unlabeled data on a directed graph* (LLUD), (Zhou et al., 2005)
    - equivalent to LGC in the case of symmetric similarities
    - assumes the input similarity graph to be strongly connected, thus considers *teleporting random walk (trw)* transition matrix

$P^\eta = \eta P + (1-\eta)P^u$ where $P = D^{-1}W$,

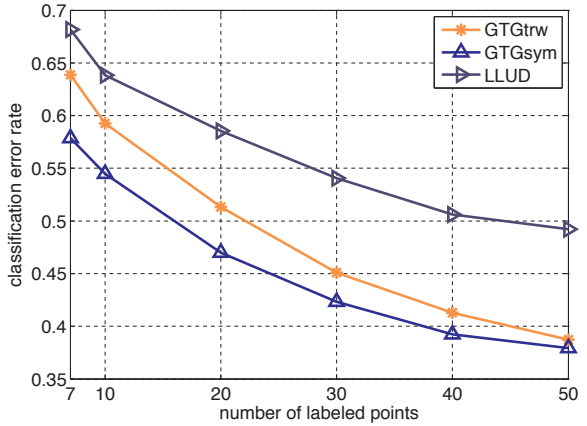
P^u - *uniform transition matrix*

- LLUD suggests a second variant for our framework (GTGtrw) where payoffs are defined in terms of P^η.

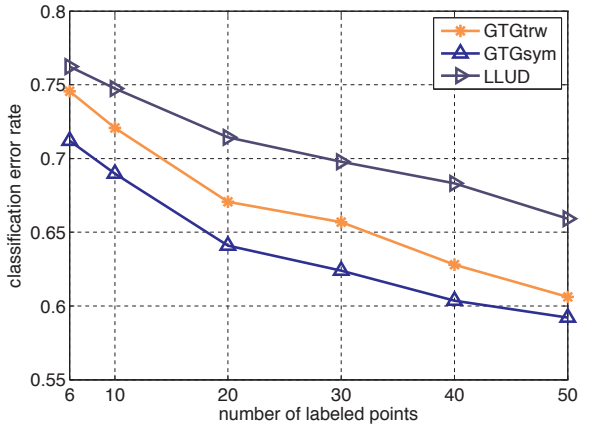
Asymmetric similarities: Results



(a) *SCOP*



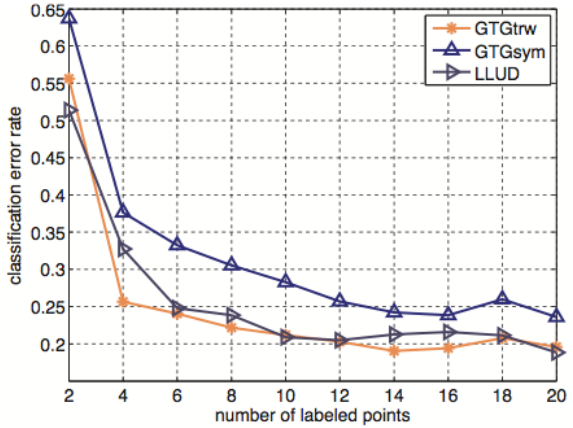
(b) *Cora*



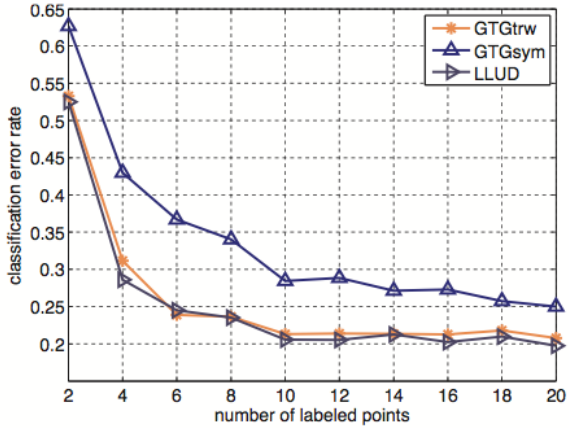
(c) *Citeseer*

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

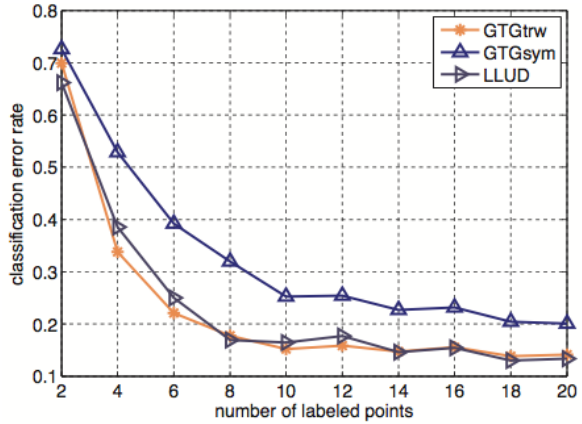
Asymmetric similarities: Results



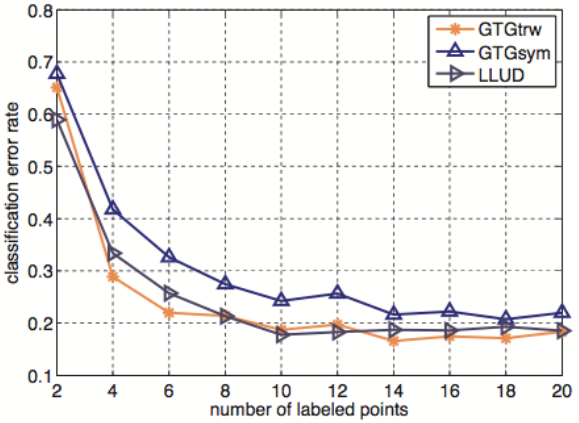
(d) *Cornell*



(e) *Texas*



(f) *Washington*



(g) *Wisconsin*

Figure: Performance comparisons on classification problems with *asymmetric* similarities (cont'd.).

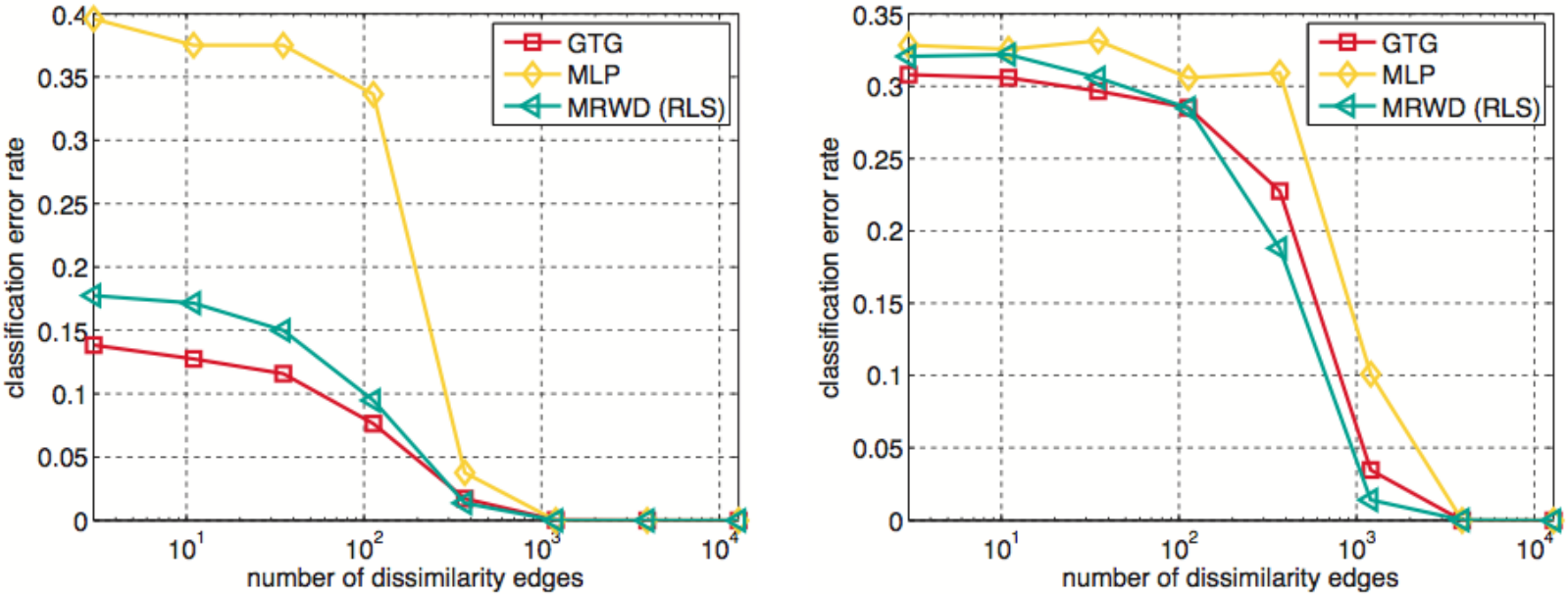
Negative similarities: Experimental Setting

- 2 data sets from UCI repository
  - Ionosphere, Diabetes

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

- Methods compared:
  - *Mixed label propagation* method (MLP) (Tong and Jin, 2007)
  - *Manifold regularization with dissimilarity* method (MRWD) (Goldberg *et al.*, 2007)
- The data sets do not originally contain negative similarities but oracle dissimilarity relations are artificially introduced by randomly sampling pairs of examples having different labels (Goldberg *et al.*, 2007)
  - $\ell = 50$
  - # of dissimilarity edges were varied between 3 and 12800
- Gaussian kernel, 19 different candidate input graphs
  - $\sigma \in \{0.01\} \cup linspace(0.05, 0.25, 5) \cup linspace(0.25, 2.5, 10) \cup \{5, 10, 20, 25\}$

Negative similarities: Results



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

Figure: The average test errors over 10 trials with randomly selected labeled examples and dissimilarity edges

Conclusions

- We addressed graph transduction from a game-theoretic view, formulating the problem as a polymatrix game.
- The proposed game-theoretic framework can cope with both negative and asymmetric similarities.
- Our results show that our approach is not only more general but also competitive with standard approaches.

Future directions

- Is there a faster way to compute Nash equilibria than using replicator dynamics?
  - See (Porter *et al.*, 2008), (Rota Bulo and Bomze,2010)
- Can we use this framework to solve even more general SSL problems?
  - Transductive learning in hypergraphs (Agarwal *et al.*, 2006), (Zhou *et al.*, 2007)
- Do we need to consider other classes of games than the class of polymatrix games?

Thanks for your attention..
Any questions?

Additional Slides

Computational Complexity

- The complexity of finding a Nash equilibrium of a graph transduction game using (**) can be given as $\mathcal{O}(kcn^2)$
  - k, the number of iterations needed to converge
  - c, the number of classes (pure strategies)
  - n, the number of data points (players)
- Experimentally, we observed that k typically grows linearly in the number of data points:

$$\mathcal{O}(kcn^2) \approx \mathcal{O}(n^3) \text{ for } c \ll n$$

  - same as the complexity of popular graph transduction methods

The problem of finding a Nash equilibrium in a polymatrix game is PPAD-complete (a subclass of NP) (Daskalakis, 2011).

PPAD: Polynomial parity argument for directed graphs

Connection to Normalized Cuts

- Ignoring the labeling constraints, (**) resembles the multi-way normalized cut criterion (Yu and Shi, 2003)

The multi-way normalized cut criterion:

$$\begin{aligned} \text{maximize} \quad & E_{NC}(Z) = \tfrac{1}{k} tr\{Z^T W Z\} \\ \text{subject to} \quad & Z^T D Z = I_k \end{aligned}$$

where $Z = X(X^T D X)^{-\frac{1}{2}}$ is the scaled version of partition matrix $X \in \{0,1\}^{n \times c},\ X\mathbf{1}_c = \mathbf{1}_n$

- Hard labeling constraints cannot be embedded into the Normalized Cuts framework in an explicit way!
- The difference in the feasible region provides robustness against noise and outliers (Pavan and Pelillo, 2007).