2012 Brazilian Symposium on Neural Networks - SBRN

Particle Competition and Cooperation to Prevent Error Propagation from Mislabeled Data in Semi-Supervised Learning

Fabricio Breve[1,2] fabricio@rc.unesp.br
Liang Zhao[2] zhao@icmc.usp.br

[1] Department of Statistics, Applied Mathematics and Computation (DEMAC), Institute of Geosciences and Exact Sciences (IGCE), São Paulo State University (UNESP), Rio Claro, SP, Brazil

[2] Department of Computer Science, Institute of Mathematics and Computer Science (ICMC), University of São Paulo (USP), São Carlos, SP, Brazil

Outline

- Learning from Imperfect Data
- The Proposed Method
- Computer Simulations
- Conclusions

Learning from Imperfect Data

- In Supervised Learning
  - Quality of the training data is very important
  - Most algorithms assume that the input label information is completely reliable
  - In practice mislabeled samples are common in data sets.

Learning from Imperfect Data

- In Semi-Supervised learning
  - Problem is more critical
    - Small subset of labeled data
    - Errors are easier to be propagated to a large portion of the data set
  - Besides its importance and vast influence on classification, it gets little attention from researchers

[4] D. K. Slonim, “Learning from imperfect data in theory and practice,” Cambridge, MA, USA, Tech. Rep., 1996.

[5] T. Krishnan, “Efficiency of learning with imperfect supervision,” *Pattern Recogn.*, vol. 21, no. 2, pp. 183–188, 1988.

[6] P. Hartono and S. Hashimoto, “Learning from imperfect data,” *Appl. Soft Comput.*, vol. 7, no. 1, pp. 353–363, 2007.

[7] M.-R. Amini and P. Gallinari, “Semi-supervised learning with an imperfect supervisor,” *Knowl. Inf. Syst.*, vol. 8, no. 4, pp. 385–413, 2005.

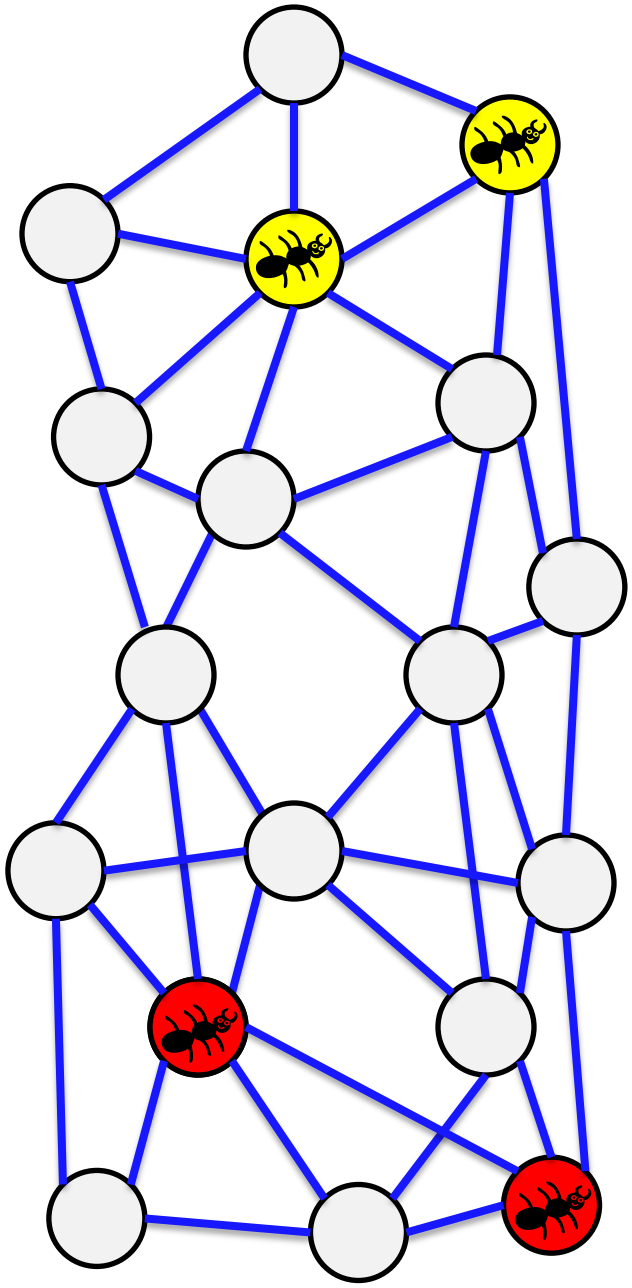
[8] ——, “Semi-supervised learning with explicit misclassification modeling,” in *IJCAI’03: Proceedings of the 18th international joint conference on Artificial intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 555–560.

Proposed Method

- Particles competition and cooperation in networks
  - Cooperation among particles representing the same team (label / class)
  - Competition for possession of nodes of the network
- Each team of particles…
  - Tries to dominate as many nodes as possible in a cooperative way
  - Prevents intrusion of particles from other teams

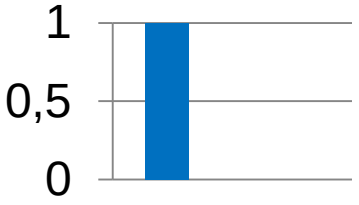
Initial Configuration

- An undirected network is generated from data by connecting each node to its *k*-nearest neighbors
  - Labeled nodes are also connected to all other nodes with the same label
- A particle is generated for each labeled node of the network
- Particles initial position are set to their corresponding nodes
- Particles with same label play for the same team



Initial Configuration

- Nodes have a domination vector
  - Labeled nodes have ownership set to their respective teams.
  - Unlabeled nodes have levels set equally for each team



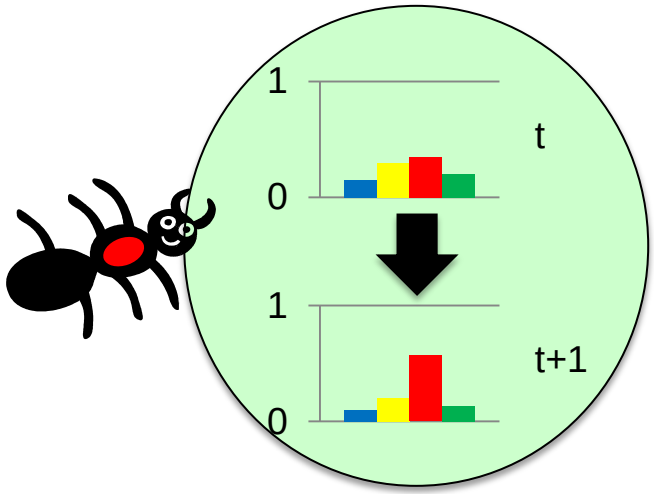
Ex: [1.00 0.00 0.00 0.00]
(4 classes, node labeled as class A)

Ex: [0.25 0.25 0.25 0.25]
(4 classes, unlabeled node)

$$v_i^{\omega_\ell}(0) = \begin{cases} 1 & \text{if } y_i = \ell \\ 0 & \text{if } y_i \neq \ell \text{ and } y_i \in L \\ \frac{1}{c} & \text{if } y_i = \emptyset \end{cases}$$

Node Dynamics

- When a particle selects a neighbor to visit:
  - It decreases the domination level of the other teams
  - It increases the domination level of its own team

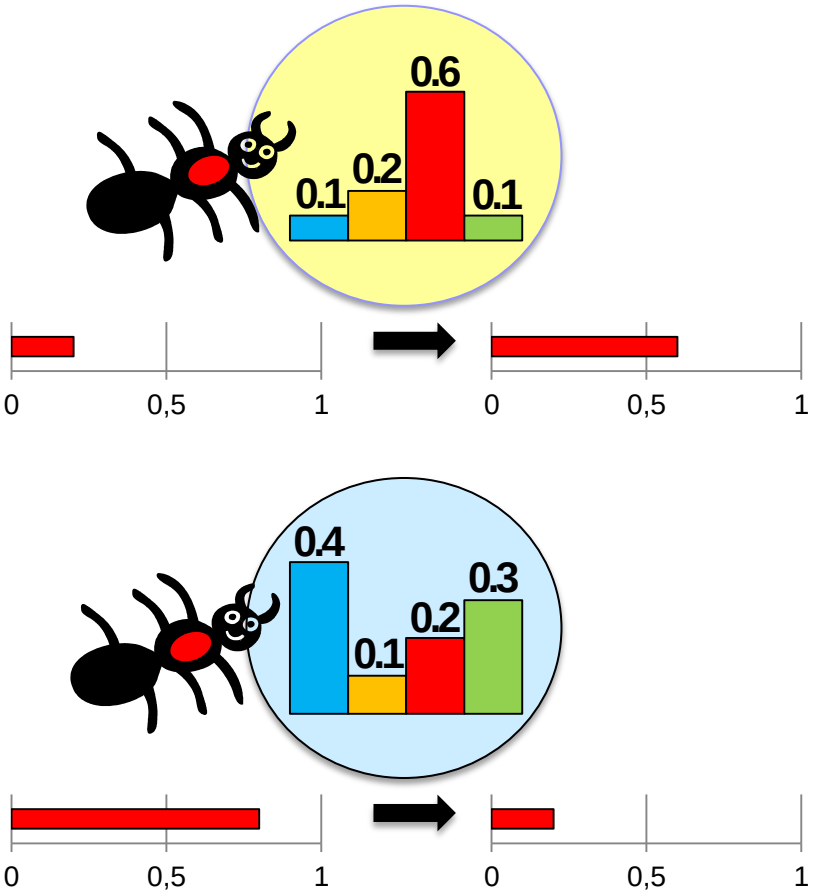


$$v_i^{\omega_\ell}(t+1) = \begin{cases} \max\{0, v_i^{\omega_\ell}(t) - \frac{0.1\rho_j^{\omega}(t)}{c-1}\} \\ \quad \ell \neq \rho_j^f \\ v_i^{\omega_\ell}(t) + \sum_{q \neq \ell} v_i^{\omega_q}(t) - v_i^{\omega_q}(t+1) \\ \quad \ell = \rho_j^f \end{cases}$$

Particle Dynamics

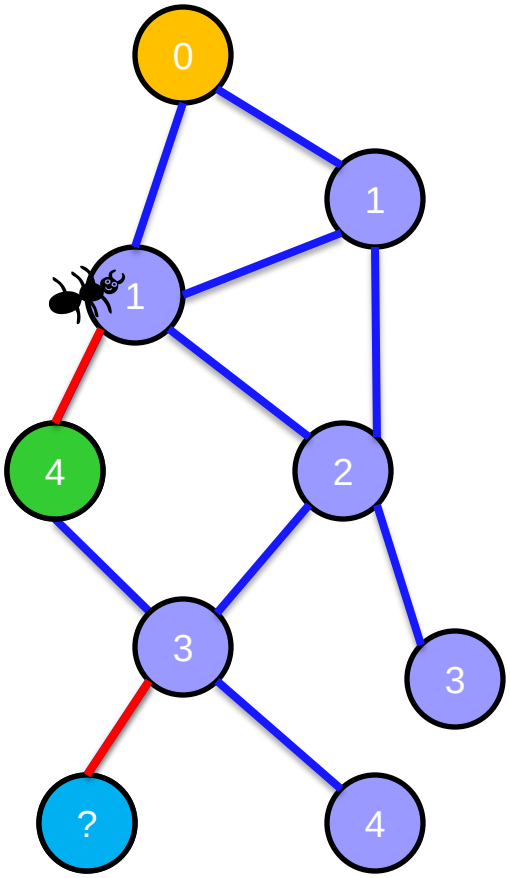
- A particle gets:
  - stronger when it selects a node being dominated by its team
  - weaker when it selects node dominated by other teams

$$\rho_j^{\omega}(t) = v_i^{\omega_\ell}(t)$$



Distance Table

- Keep the particle aware of how far it is from the closest labeled node of its team (class)
  - Prevents the particle from losing all its strength when walking into enemies neighborhoods
  - Keep them around to protect their own neighborhood.
- Updated dynamically with local information
  - Does not require any prior calculation

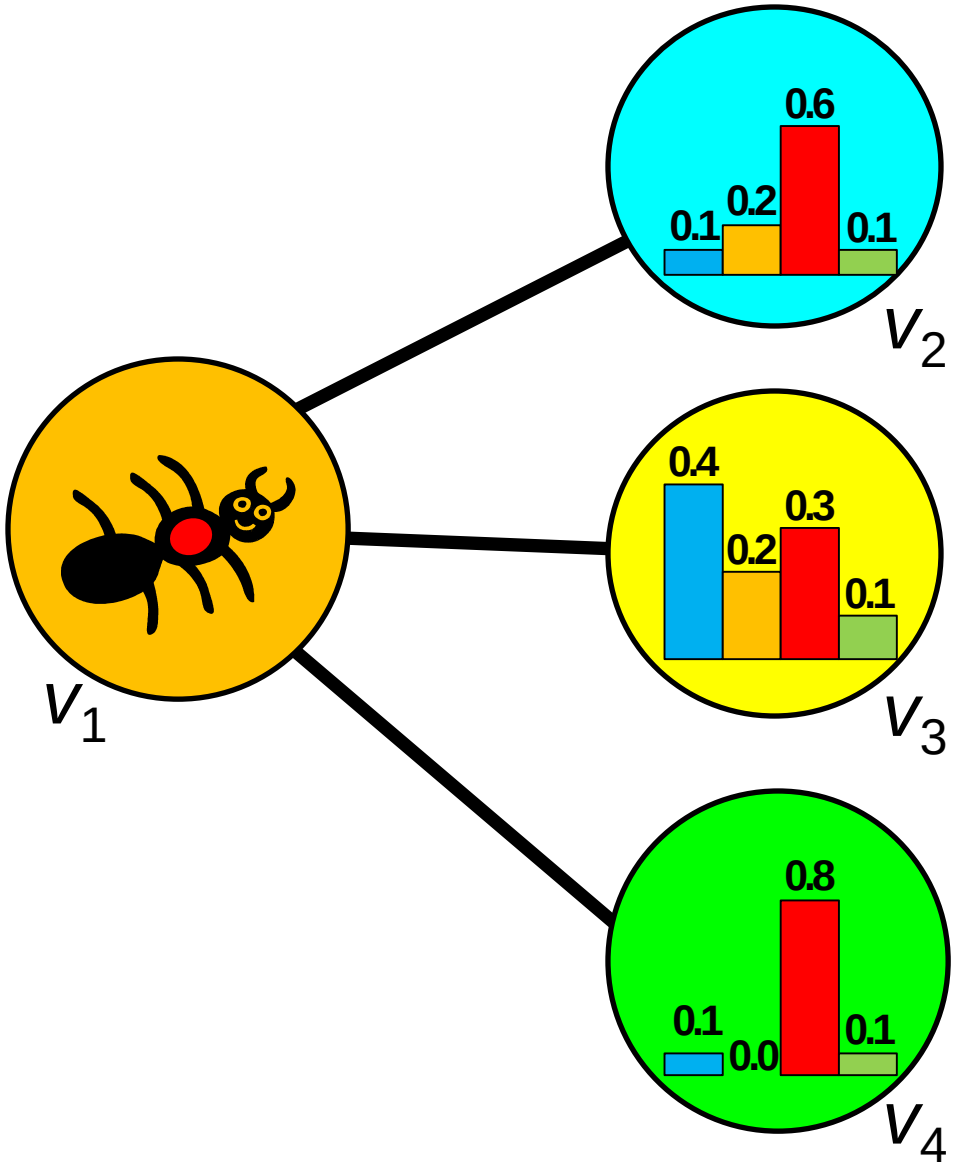


$$\rho_j^{d_k}(t+1) = \begin{cases} \rho_j^{d_i}(t) + 1 & \text{if } \rho_j^{d_i}(t) + 1 < \rho_j^{d_k}(t) \\ \rho_j^{d_k}(t) & \text{otherwise} \end{cases}$$

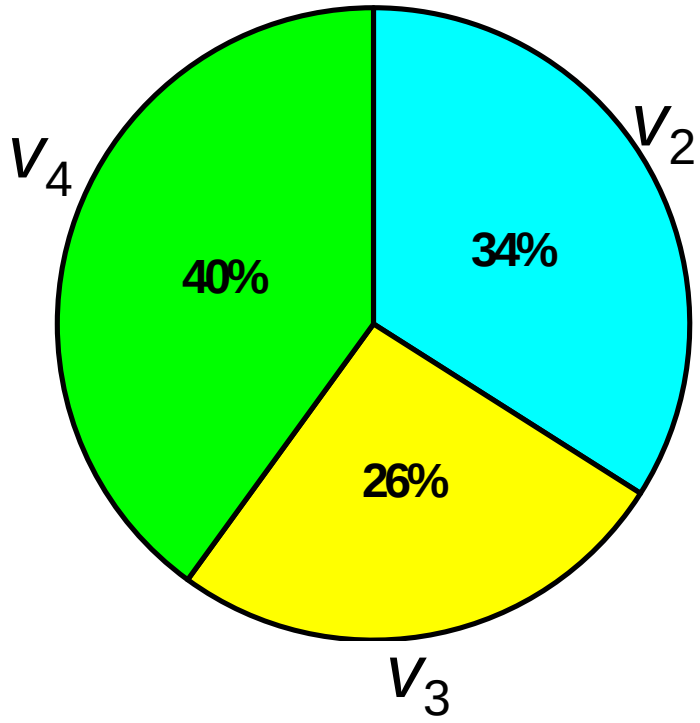
Particles Walk

- Random-greedy walk
  - The particle will prefer visiting nodes that its team already dominates and nodes that are closer to the labeled nodes of its team (class)

$$p(v_i|\rho_j) = 0.5 \left(\frac{W_{qi}}{\sum_{\mu=1}^{n} W_{q\mu}} + \frac{W_{qi} v_i^{\omega_\ell} \frac{1}{(1+\rho_j^{d_i})^2}}{\sum_{\mu=1}^{n} W_{q\mu} v_i^{\omega_\ell} \frac{1}{(1+\rho_j^{d_i})^2}} \right)$$

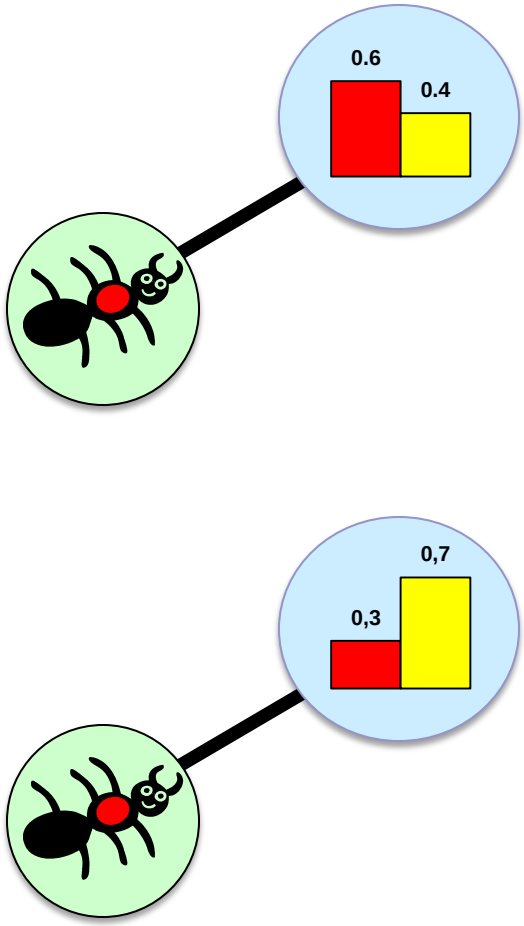


Moving Probabilities



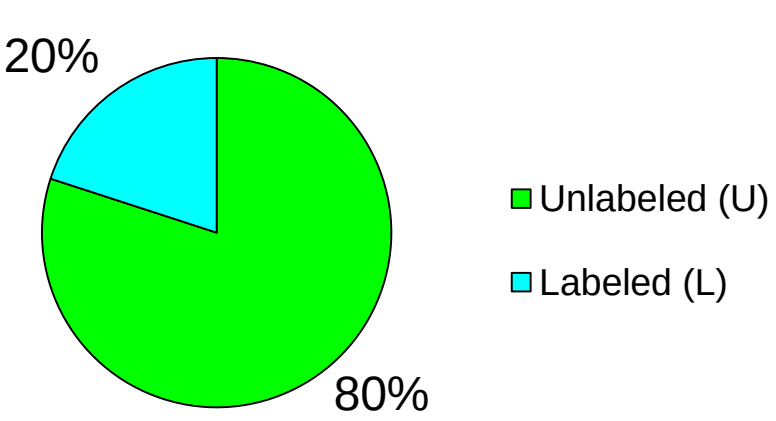
Particles Walk

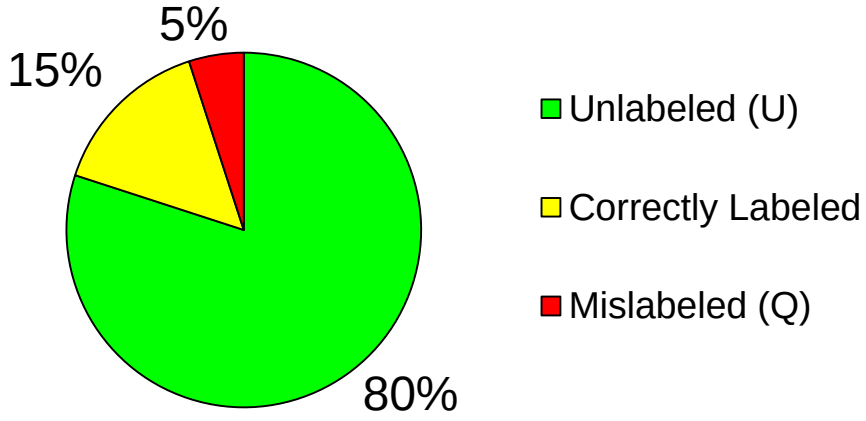
- Shocks
  - A particle really visits the selected node only if the domination level of its team is higher than others;
  - otherwise, a shock happens and the particle stays at the current node until next iteration.

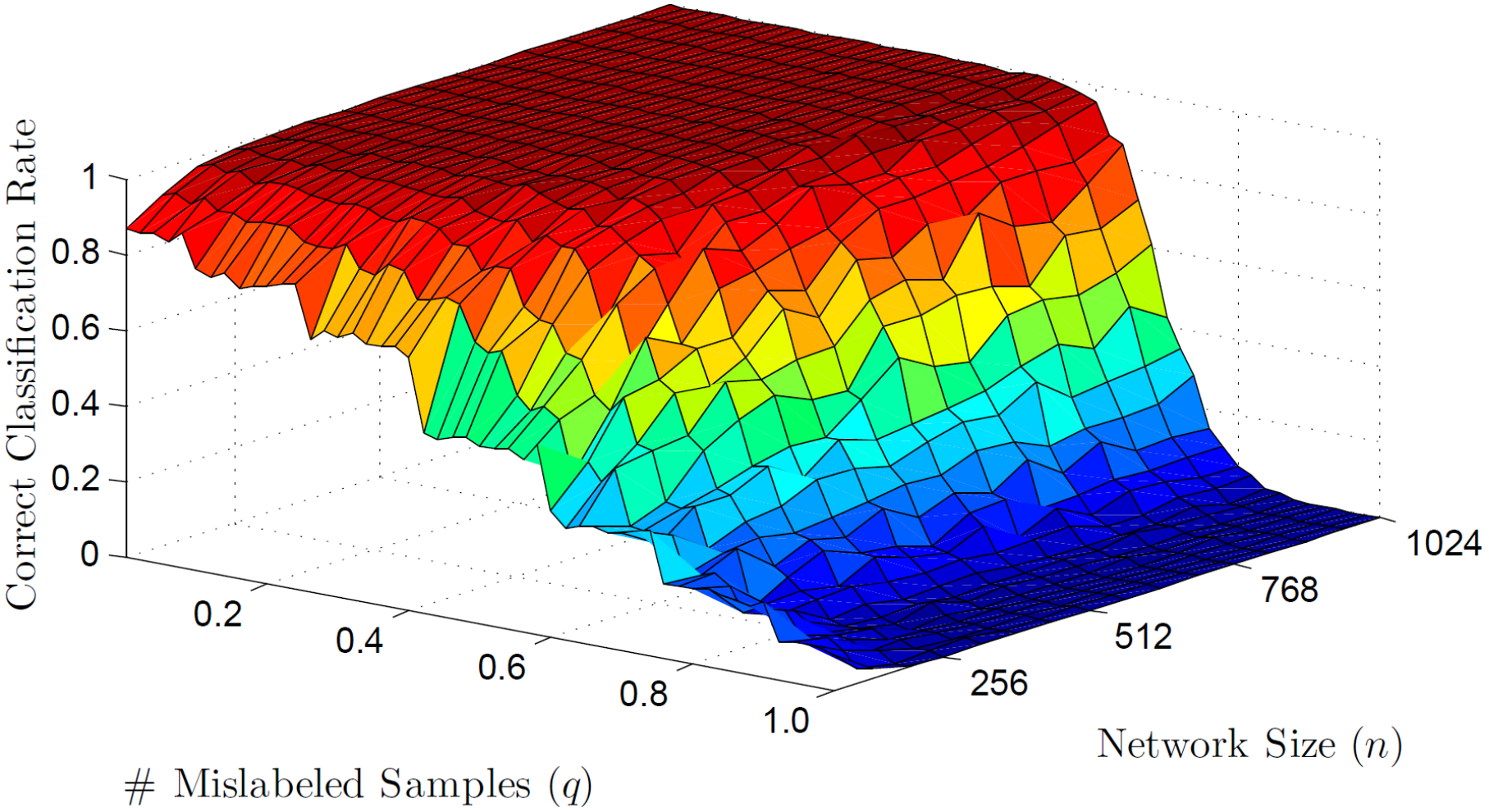


Computer Simulations

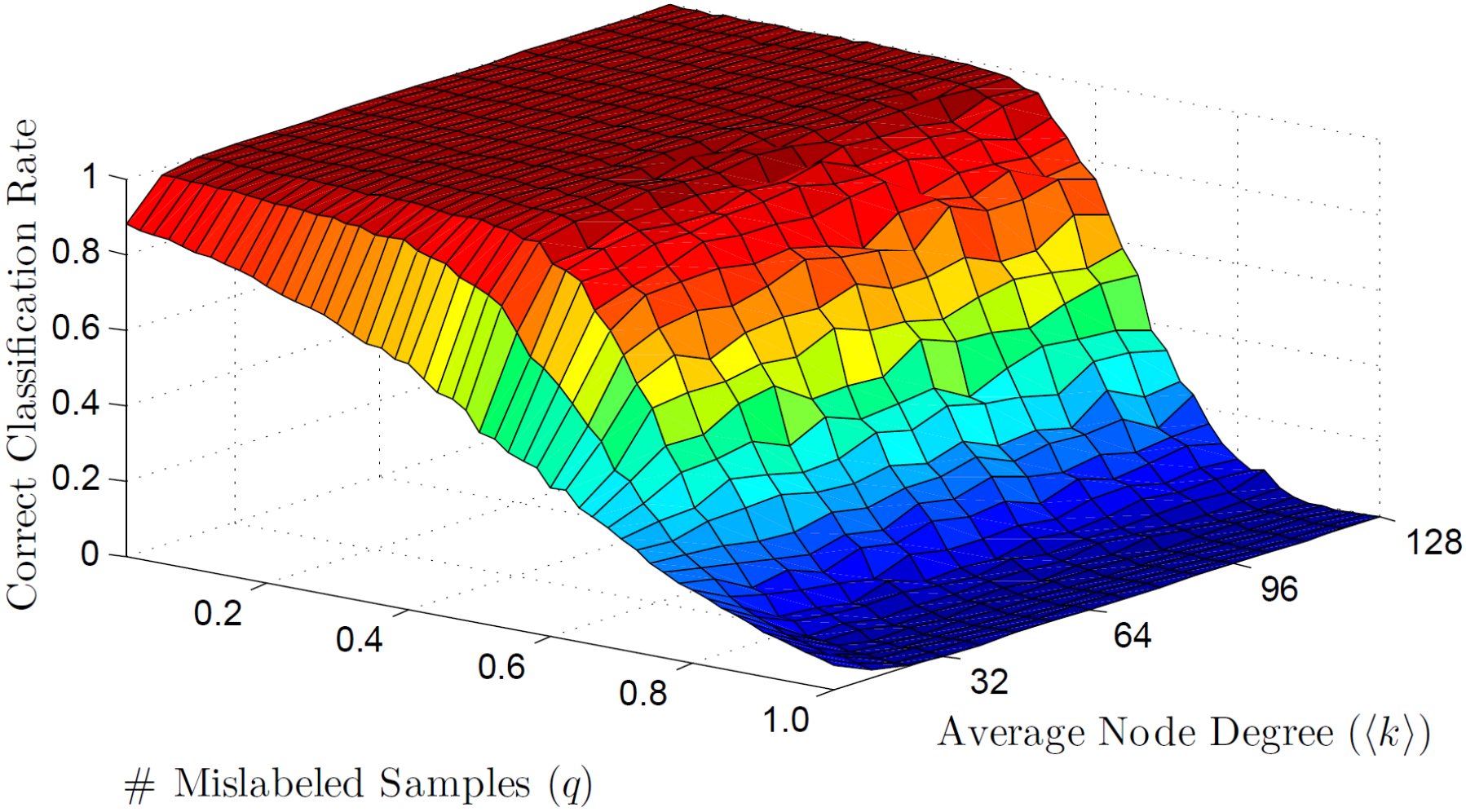
- Network are generated with:
  - Different sizes and average node degrees
  - Elements divided into 4 classes
  - 25% of the edges are connecting different classes nodes
  - Set of nodes *N*
  - Labeled subset $L \subset N$
  - Mislabeled subset $Q \subset L \subset N$



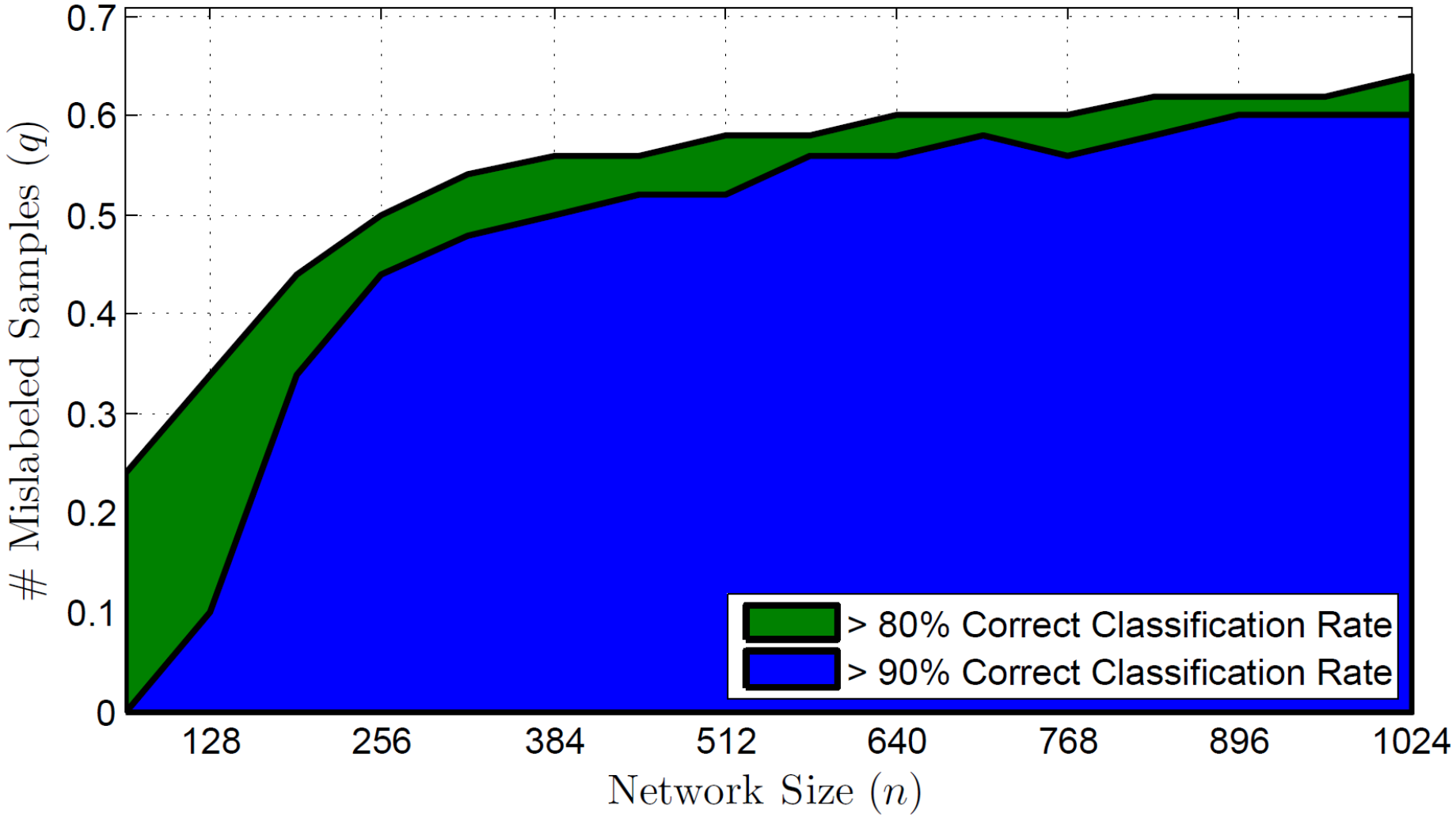




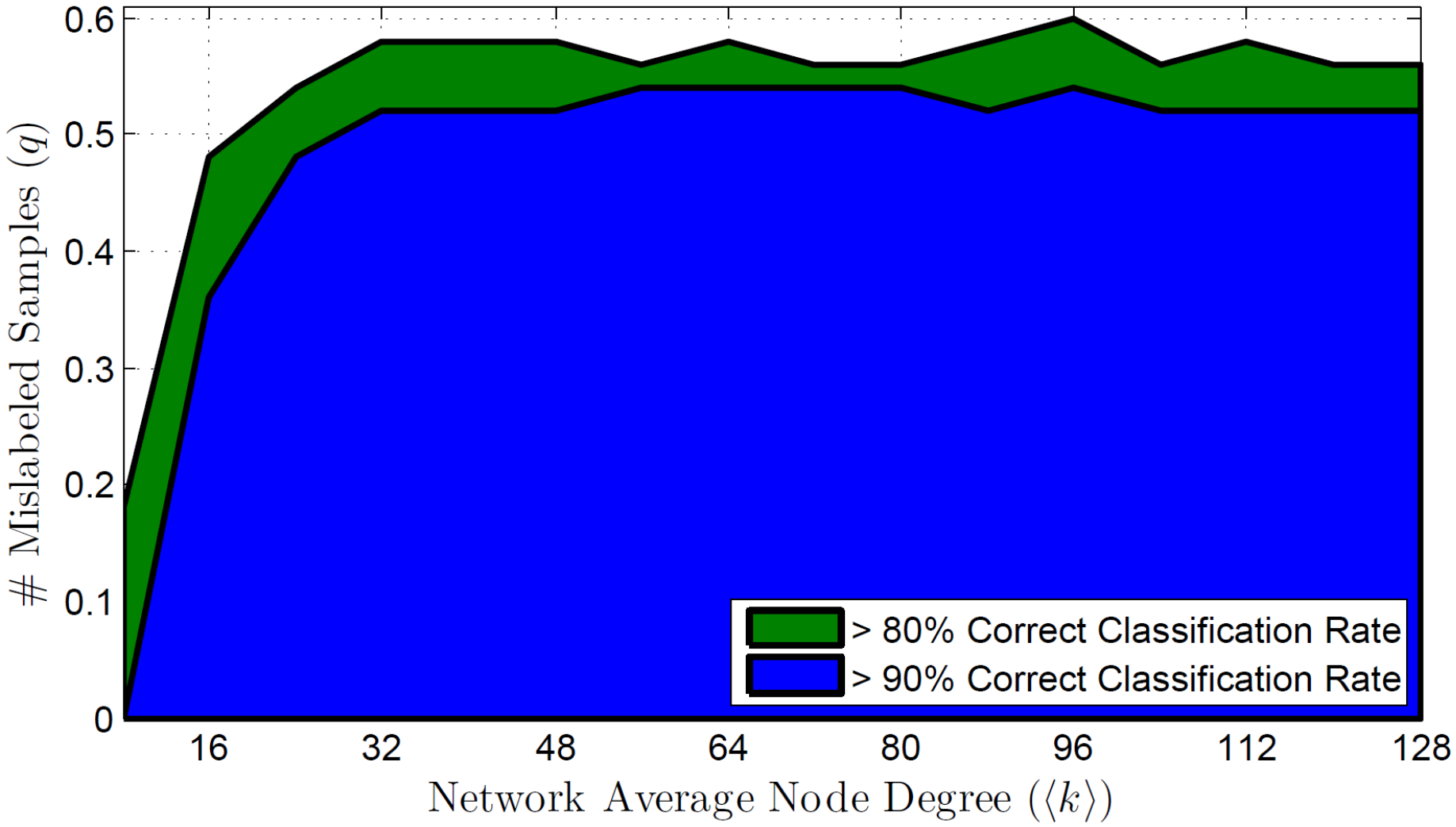
Correct Classification Rate with different network sizes and mislabeled subset sizes, $\langle k \rangle = n/8$, $l = n/0.1$



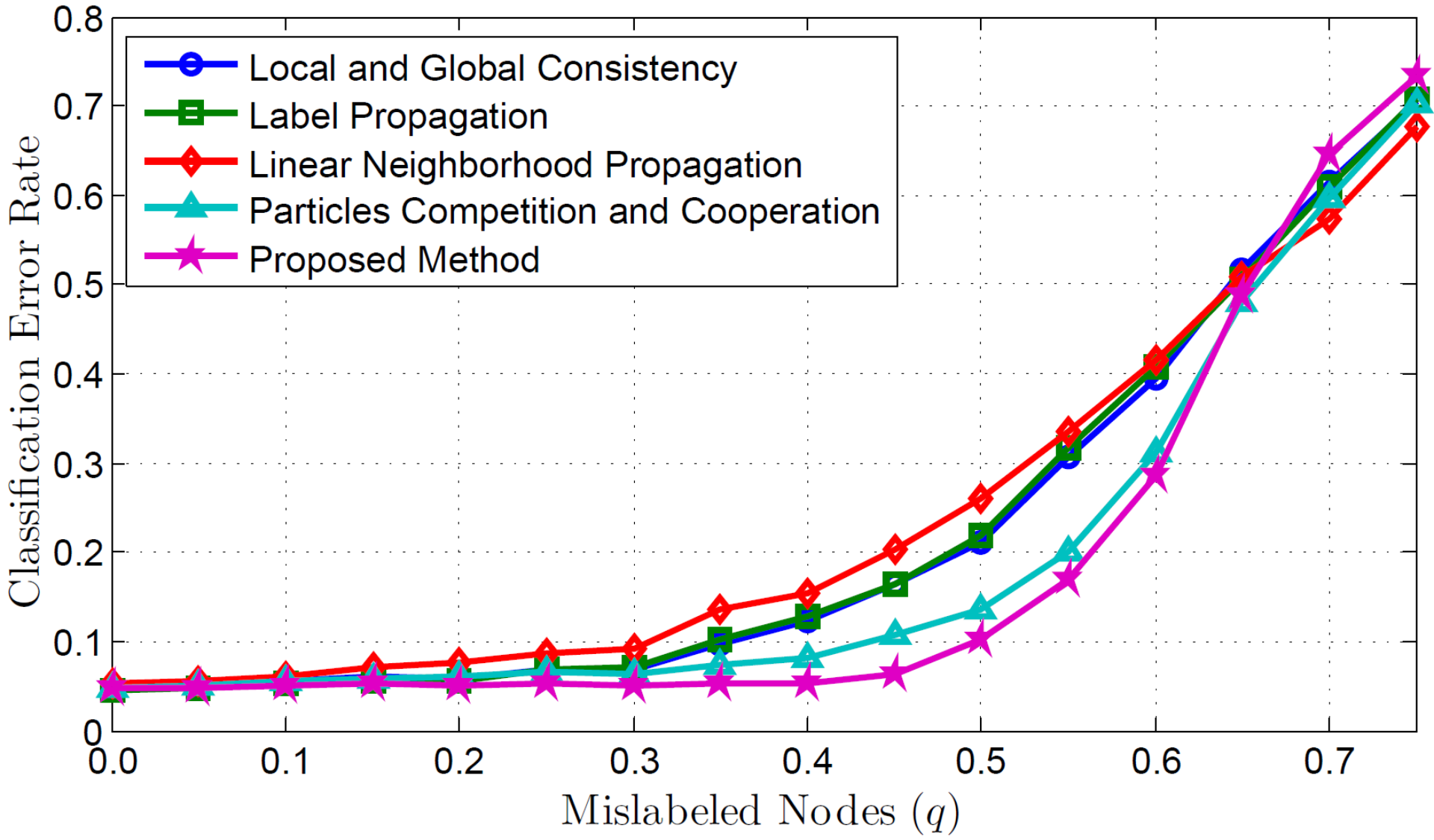
Correct Classification Rate with different average node degrees and mislabeled subset sizes, n = 512, l = 64.



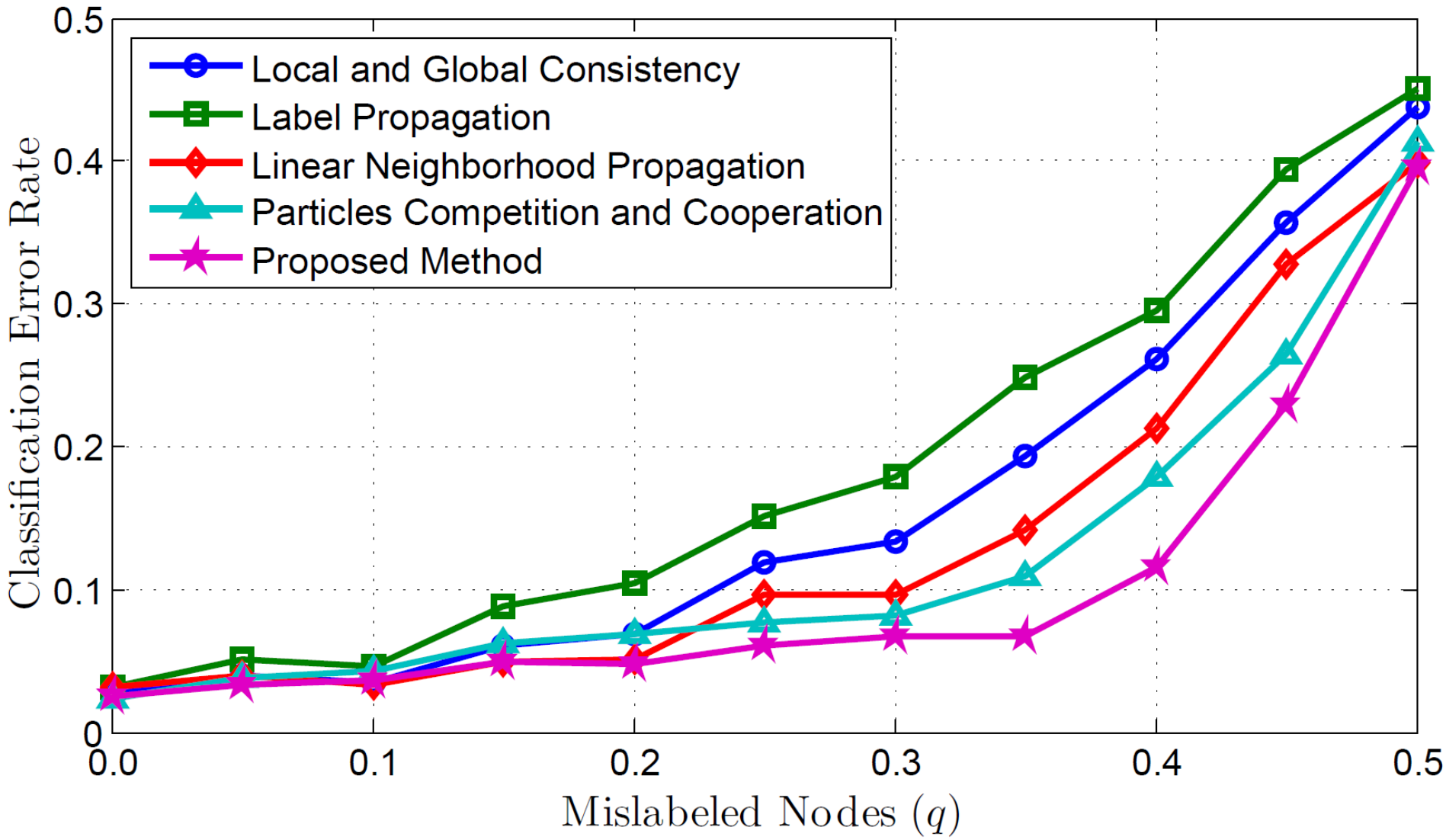
Maximum mislabeled subset size for 80% and 90% of correct classification rate with different network sizes, $\langle k \rangle = n/8$, $z_{out}/\langle k \rangle = 0.25$, $l/n = 0.1$



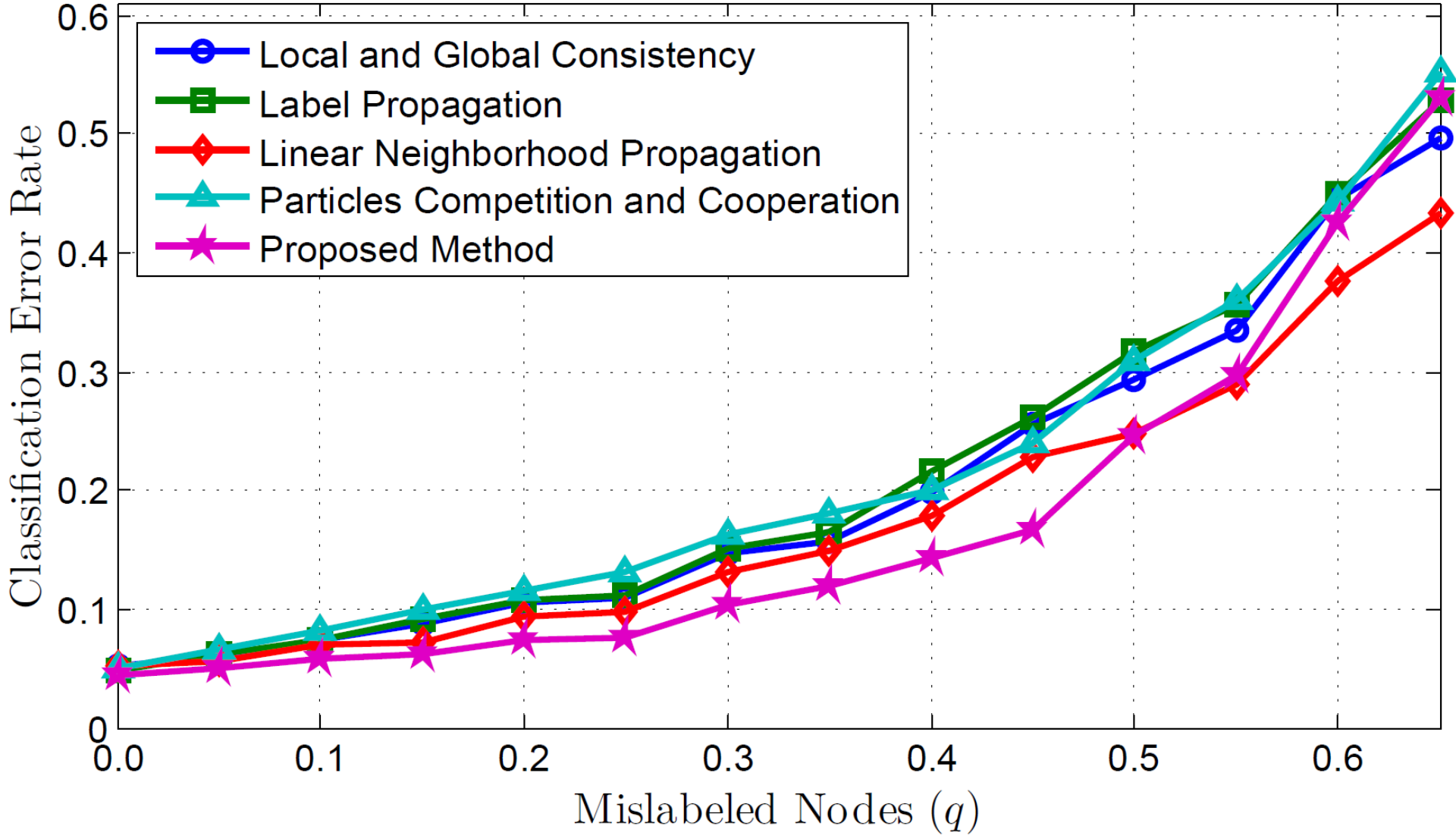
Maximum mislabeled subset size for 80% and 90% of correct classification rate with different network average node degree (⟨*k*⟩), *n* = 512, *l*/*n* = 0.1



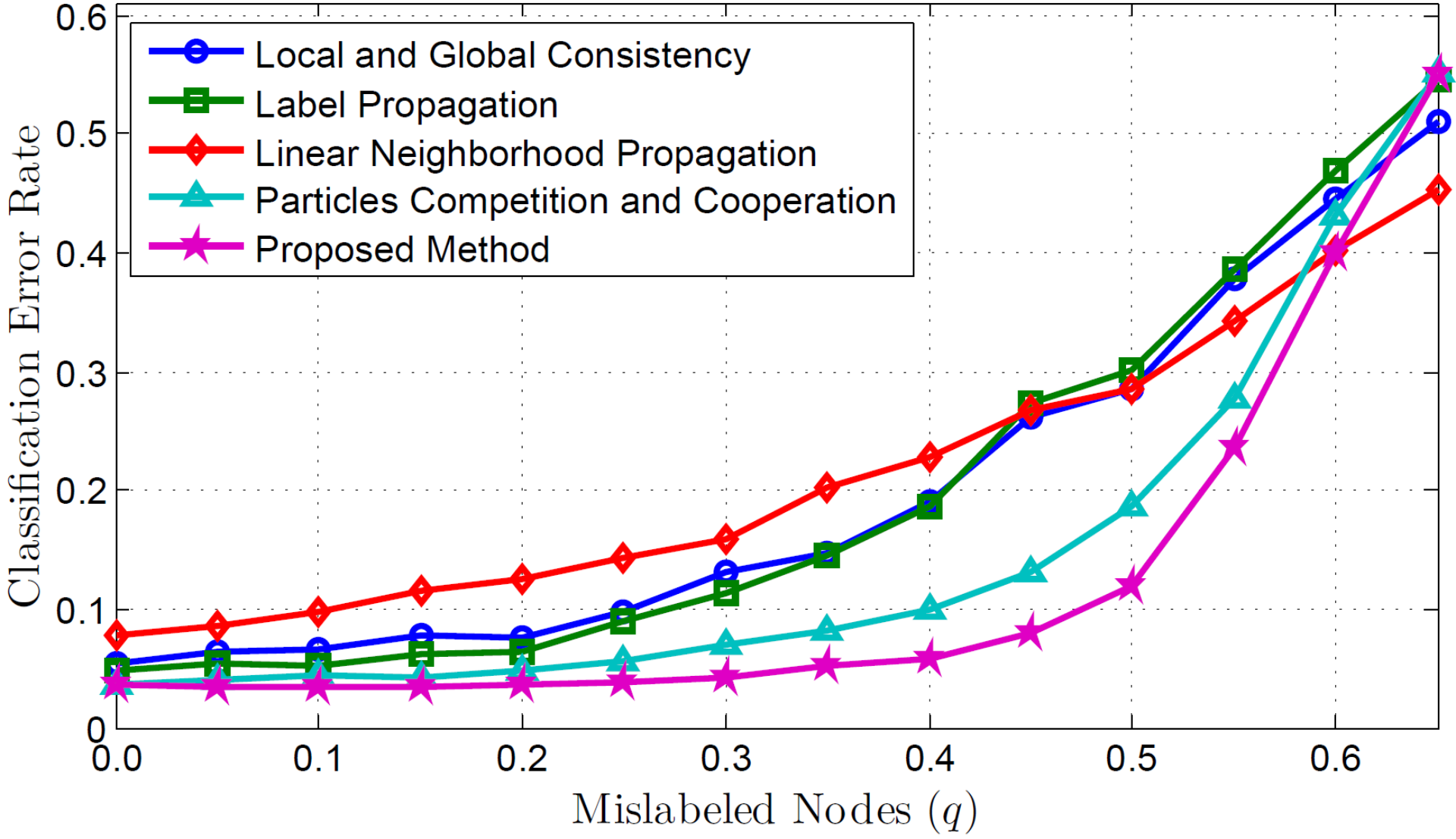
Classification error rate in a network with 4 normally distributed classes with different mislabeled subset size



Classification error rate in the Digit1 data set with different mislabeled subset size



Classification error rate in the Iris data set with different mislabeled subset size
40 labeled samples



Classification error rate in the Wine data set with different mislabeled subset size
40 labeled samples

Conclusions

- New biologically inspired method for semi-supervised classification
  - Specifically designed to handle data sets with mislabeled subsets
    - *A mislabeled node may have its label changed when the team which has its correct label first dominates the nodes around it, then attacks it, and finally takes it over, thus stopping wrong label propagation from that node*

Conclusions

- Results analysis indicate the presence of critical points in the performance curve as the mislabeled samples subset grows.
  - Related to the network size and average node degree.
- Proposed algorithm
  - Shows robustness in the presence of mislabeled data.
  - Performed better than other representative graph-based semi-supervised methods when applied to artificial and real-world data sets with mislabeled samples.

Future Work

- Expand the analysis to cover the impact of other networks measures in the algorithm performance
- Expand the comparison to include more and larger data sets with mislabeled nodes

Acknowledgements

- This work was supported by:
  - State of São Paulo Research Foundation (FAPESP)
  - Brazilian National Council of Technological and Scientific Development (CNPq)
  - Foundation for the Development of Unesp (Fundunesp)

2012 Brazilian Symposium on Neural Networks - SBRN

Particle Competition and Cooperation to Prevent Error Propagation from Mislabeled Data in Semi-Supervised Learning

Fabricio Breve[1,2] fabricio@rc.unesp.br
Liang Zhao[2] zhao@icmc.usp.br

[1] Department of Statistics, Applied Mathematics and Computation (DEMAC), Institute of Geosciences and Exact Sciences (IGCE), São Paulo State University (UNESP), Rio Claro, SP, Brazil

[2] Department of Computer Science, Institute of Mathematics and Computer Science (ICMC), University of São Paulo (USP), São Carlos, SP, Brazil