Data Mining & Machine Learning

ARIZONA STATE UNIVERSITY

Toward Relational Learning with Misinformation

Liang Wu*, Jundong Li*, Fred Morstatter+, Huan Liu*

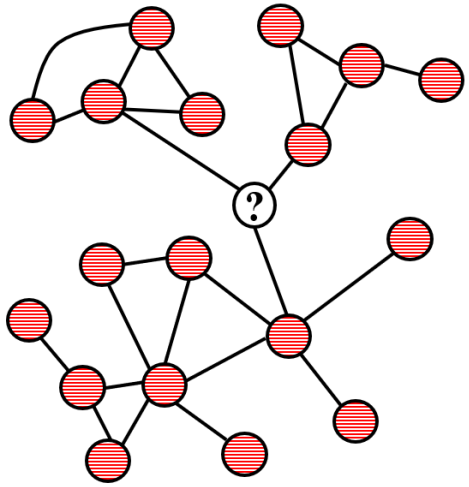
*Arizona State University

+University of Southern California

{wuliang, jundongl, huanliu}@asu.edu, morstatt@usc.edu

Classification in Social Media

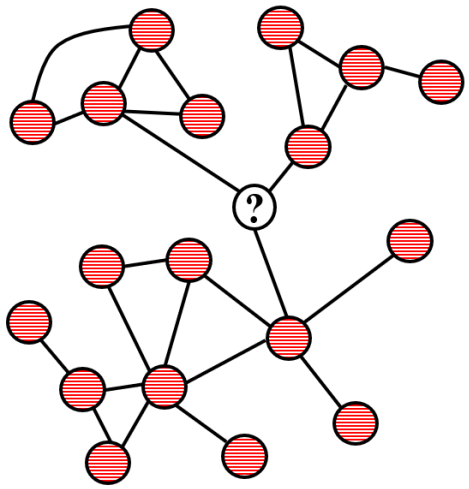
- Relational learning aims to **classify linked nodes** in a graph (social networks)



- Task: Classification
- Feature: Attributes, Links

Classification in Social Media: Our Task

- Relational learning aims to **classify linked nodes** in a graph (social networks)



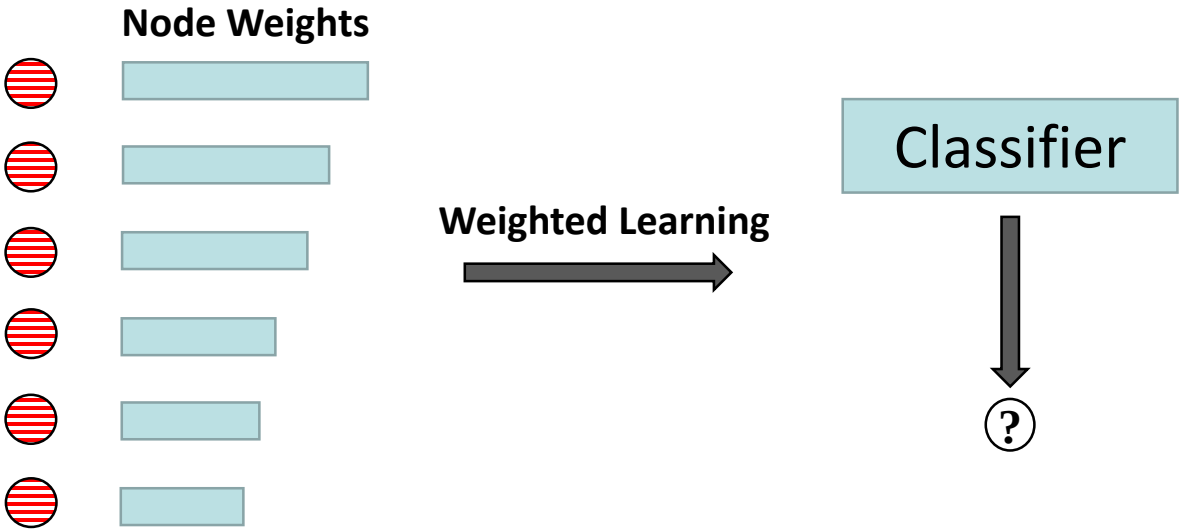
- Task: Classification
- Feature: Attributes, Links
- Challenge: Data is Inaccurate

Social Media Data is Inaccurate and Noisy

- Attacks of content polluters
  - Node attributes cannot reveal the identity
- Colloquial language of regular users
  - Misinformation, inaccurate data

Classification with Noisy Data

- Weighting Nodes



- Anomalous points are lower weighted
  - Larger loss leads to smaller weights

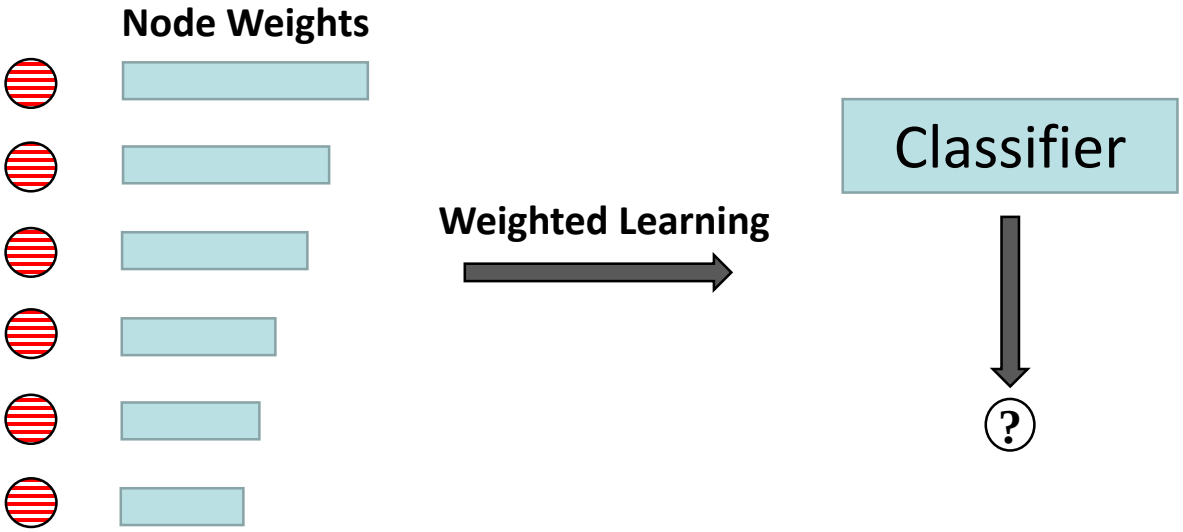
Classification with Noisy Social Media Data

- Attacks of content polluters
  - Node attributes cannot reveal the identity

- Colloquial language of regular users
  - Misinformation, inaccurate data

Robust Classification with Network Information

- Weighting Nodes with Centrality



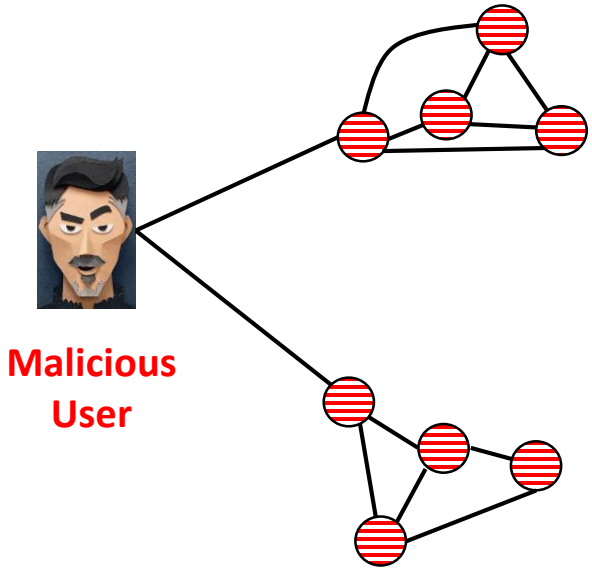
- Authoritative points are higher weighted
  - ~~Larger loss leads to smaller weights~~
  - Larger **centrality** leads to **higher** weights

Denoising with Social Networks?

- Links can be noisy

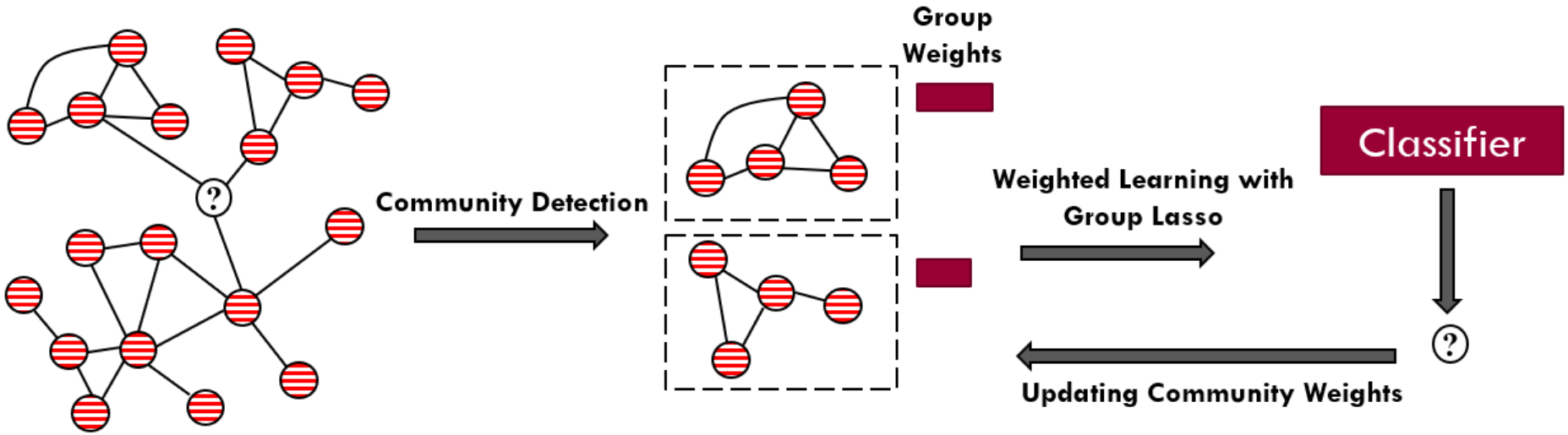
- Obtaining all links (complete graph) is difficult

Community Structures are More Robust



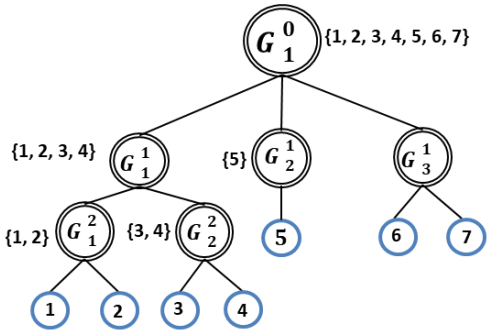
Community Structures are More Robust

Denoise with Community Structures

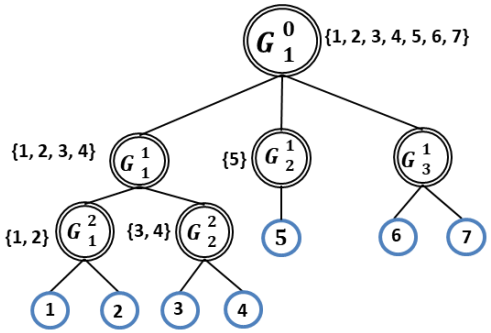


Estimating Communities with Network Structures → Weighting Nodes with Community Structures → Training Classifiers Updating Group Weights

Community Candidate Generation + Community Selection



Community Candidate Generation + Community Selection



$$\min_{\mathbf{w},\mathbf{c}} \sum_{i=1}^{N} c_i (\mathbf{x_i w} - \mathbf{y}_i)^2 + \lambda_1 ||\mathbf{w}||_2^2 + \lambda_2 \sum_{i=0}^{d} \sum_{j=1}^{n_i} ||\mathbf{c}_{G_j^i}||_2$$

avoid overfitting

group Lasso

$$\text{Subject to } \sum_i c_i = K$$

L_1 norm on the inter-group level

L_2 norm on the intra-group level

d: depth of hierarchy of Louvain method

n_i: number of groups on layer i

$\mathbf{c}_{G_j^i}$: nodes of group j on layer i

Optimization

Optimize **w**

$$\min_{\mathbf{w}} \sum_{i=1}^{m} c_i \left(\mathbf{x_i w} - \mathbf{yi}\right)^2 + \lambda_1 ||\mathbf{w}||_2^2$$

Optimize **c**

$$\min_{\mathbf{w},\mathbf{c}} \sum_{i=1}^{m} c_i \left(t_i\right) + \lambda_2 \sum_{i=0}^{d} \sum_{j=1}^{n_i} ||\mathbf{c}_{G_j^i}||_2$$

$$\text{Subject to } \sum_i c_i = 1$$

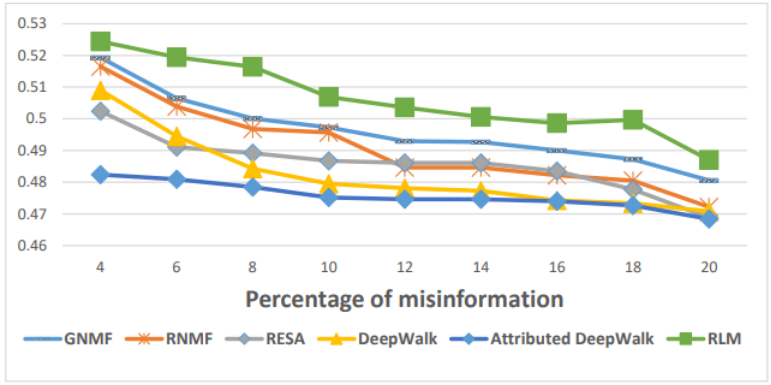
Evaluation

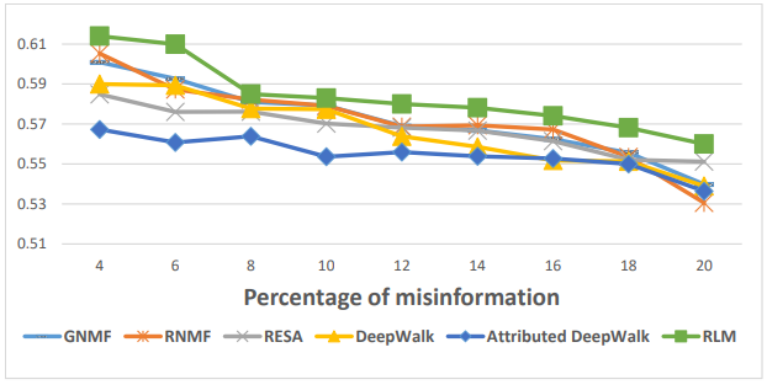
Datasets

Dataset	#Instances	#Labels	#Features
Blog Catalog	5,198	6	8,189
Flickr	7,575	9	12,047

Results

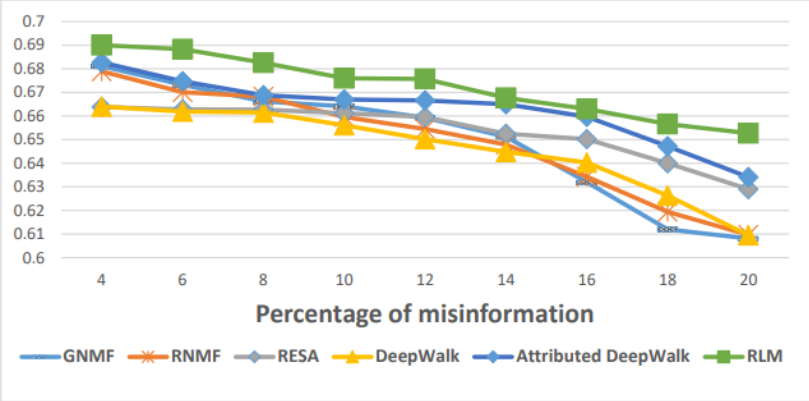
Macro- and Micro-average of F_1-measures with increasing ratio of misinformation

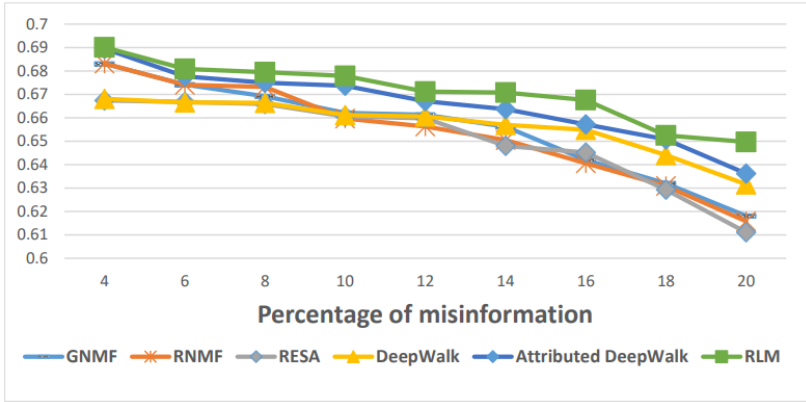




Flickr

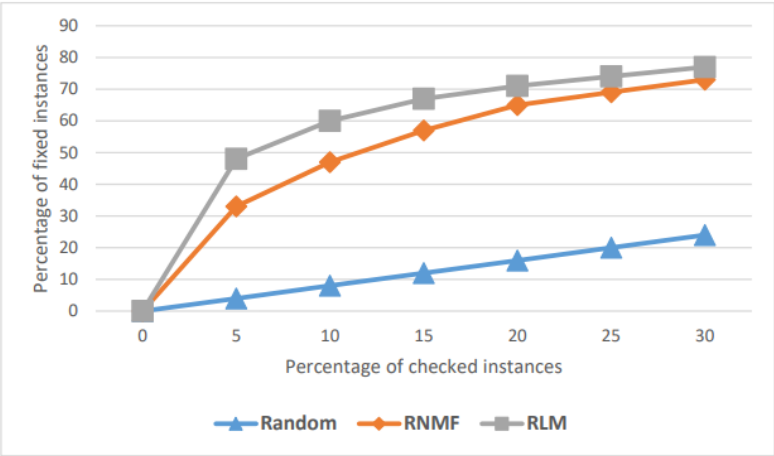
More Results



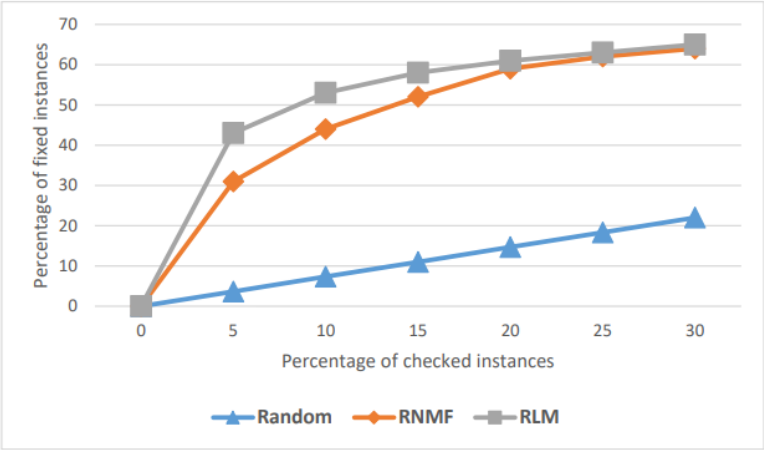


BlogCatalog

Effectiveness of identifying mislabeled instances



BlogCatalog



Flickr

Conclusions

- A supervised learning method with inaccurate networked data
  - Focusing on community structures instead of links
  - Can be integrated to other algorithms
  - Efficient to solve

Liang Wu*, Jundong Li*, Fred Morstatter+, Huan Liu*

*Arizona State University

+University of Southern California

{wuliang, jundongl, huanliu}@asu.edu, morstatt@usc.edu