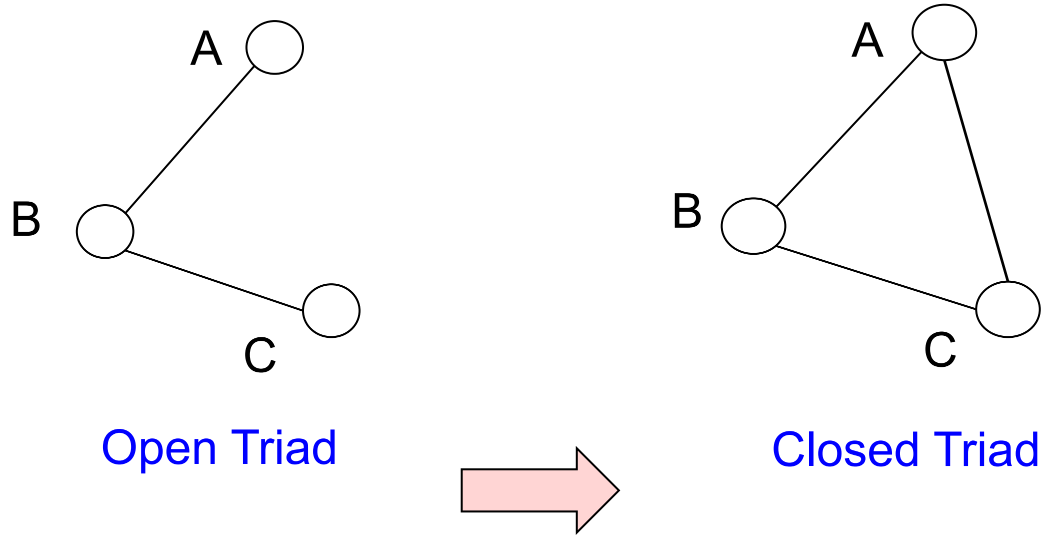
Uncovering the Formation of Triadic Closure in Social Networks

Zhanpeng Fang and Jie Tang
Tsinghua University

Triangle 'Laws'

- Triangle is one of most basic human groups in social networks
  - Friends of friends are friends



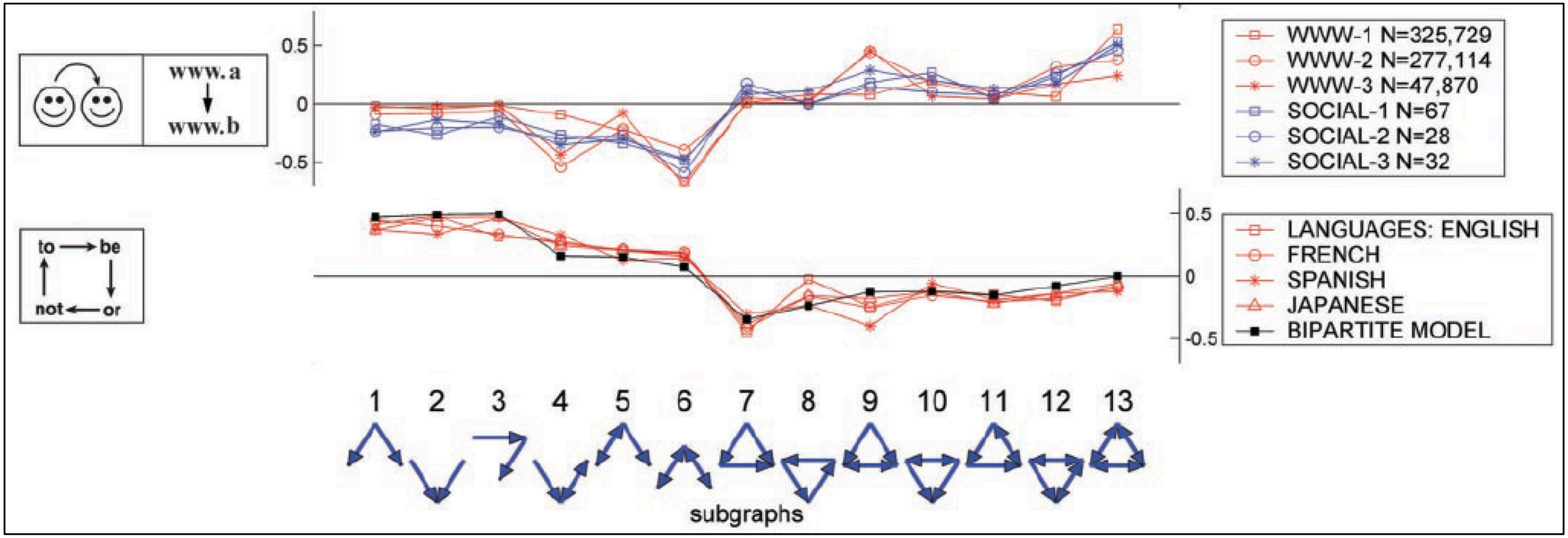
Open Triad

Closed Triad

Triadic Closure Process

Triadic Closure

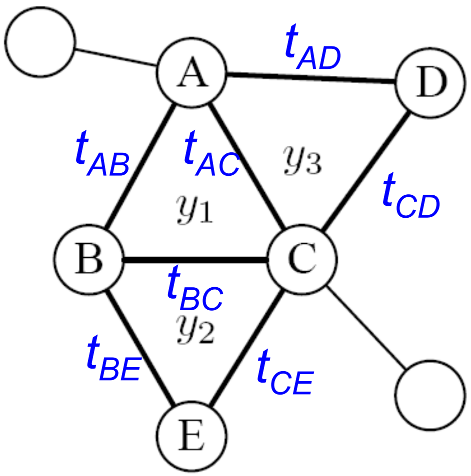
- Uncovering the mechanism underlying the triadic closure process can benefit many applications
  - Classify different types of networks[1]
  - Explain the evolution of social communities[2]



[1] Milo, Ron, et al. "Superfamilies of evolved and designed networks." *Science* (2004)
[2] Kossinets, Gueorgi, and Duncan J. Watts. "Empirical analysis of an evolving social network." Science (2006)

Decoding Triadic Closures

- Goal: Uncovering how each closed triad was formed step by step



$$y_1=(t_{AB} \succ t_{BC} \succ t_{AC})$$

$$y_2=(t_{BE} \succ t_{BC} \succ t_{CE})$$

  – Challenge: Target space is large and continuous.

- Focus on detecting the partial order of the formation time of the three links in a closed triad

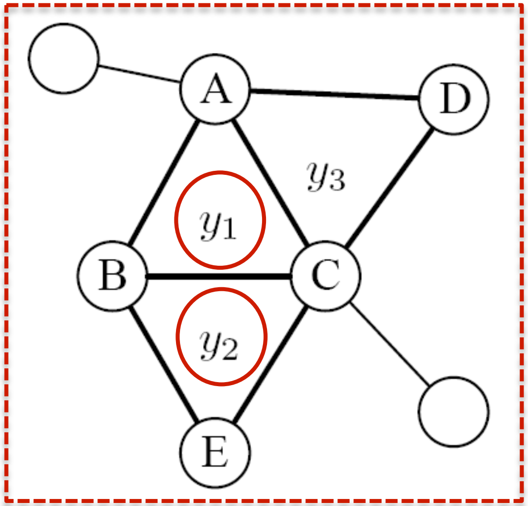
Problem Definition – Decoding Triadic Closure

Input: social network $G=(V,E)$

A small set of labeled results Y^L

A large set of unlabeled triads $\{\triangle\}^U$

Output: $f : (\{\triangle\}^U | G, Y^L) \rightarrow Y^U$



$y_1=(t_{AB} \succ t_{BC} \succ t_{AC})$

$y_2=(t_{BE} \succ t_{BC} \succ t_{CE})$

$y_3 = ?$

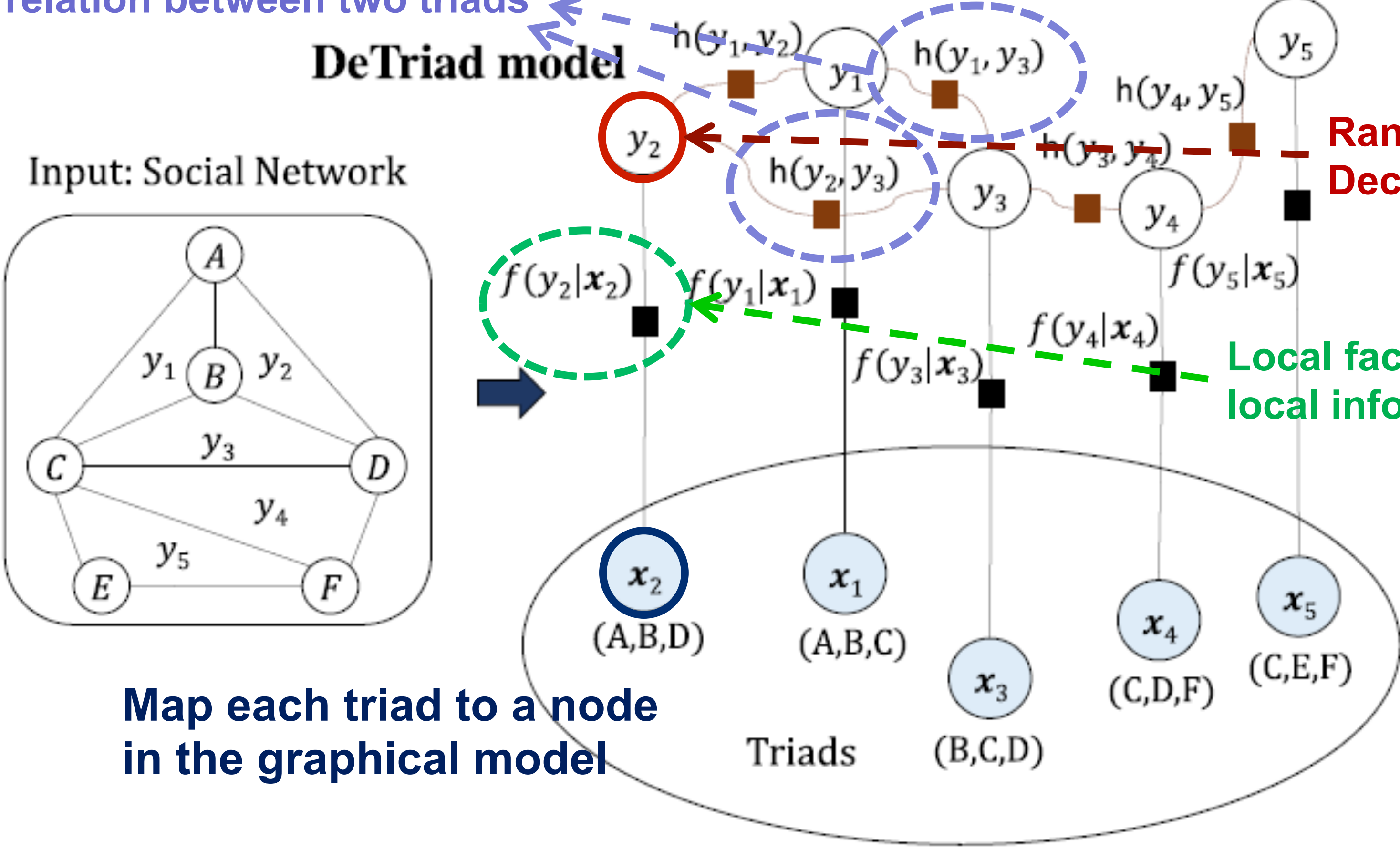
$Y^L=\{y_1, y_2\}$

$\{\triangle\}^U=\{\triangle ACD\}$

$Y^U=\{y_3\}$

DeTriad—the proposed Model

Correlation factor h(): Modeling correlation between two triads



Random variable Y: Decoding result

Local factor f(): Modeling local information

Map each triad to a node in the graphical model

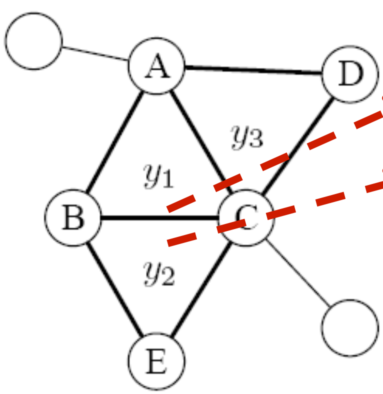
Joint Distribution: $P(Y|\boldsymbol{X},G) = \prod_{\Delta_i} f(y_i|\boldsymbol{x}_i) \prod_{i\sim j} h(y_i, y_j)$

DeTriad Model (cont')

Joint Distribution:
$$P(Y|\boldsymbol{X},G) = \prod_{\Delta_i} \boxed{f(y_i|\boldsymbol{x}_i)} \prod_{i\sim j} \boxed{h(y_i,y_j)}$$

Local Factor:
$$\boxed{f(y_i|\boldsymbol{x}_i) = \frac{1}{Z_1}\exp\{\sum_{k=1}^{d}\alpha_k f_k(x_{ik},y_i)\}}$$

Correlation Factor:
$$\boxed{h(y_i,y_j) = \frac{1}{Z_2}\exp\{\sum_{k}\mu_k h_k(y_i,y_j)\}}$$



K_1: Rank of BC in △ABC

K_2: Rank of BC in △BCE

$$h(y_i,y_j) = \frac{1}{Z_3}\exp\{\sum_{k}\mu_k \cdot I_k(y_i,y_j)\}$$

Synchronous method: Consider $K_1 = K_2$

$$h(y_i,y_j) = \frac{1}{Z_4}\exp\{\sum_{k_i,k_j}\mu_{k_i,k_j} \cdot I_{k_i,k_j}(y_i,y_j)\}$$

Asynchronous method: Consider all possible K_1, K_2

DeTriad Model (cont')

- Objective function: $\mathcal{O}(\theta) = \log P(Y^L|\mathbf{X}, G) = \log \sum_{Y|Y^L} P(Y|\mathbf{X}, G)$

Incorporate partial labeled information

$$\boxed{\begin{aligned} &= \log \sum_{Y|Y^L} \{\sum_{\Delta_i} \sum_{k=1}^{d} \alpha_k f_k(x_{ik}, y_i) + \sum_{i \sim j} \sum_{k} \mu_k h_k(y_i, y_j)\} \\ &- \log \sum_{Y} \{\sum_{\Delta_i} \sum_{k=1}^{d} \alpha_k f_k(x_{ik}, y_i) + \sum_{i \sim j} \sum_{k} \mu_k h_k(y_i, y_j)\} \end{aligned}}$$

- Model learning: Gradient descent

$$\frac{\partial \mathcal{O}(\boldsymbol{\theta})}{\partial \mu_k} = \mathbf{E}_{P_{\mu_k}(y_i, y_j | Y^L, \mathbf{X}, G)}[h_k(y_i, y_j)] - \mathbf{E}_{P_{\mu_k}(y_i, y_j | \mathbf{X}, G)}[h_k(y_i, y_j)]$$

- Decoding for triad $\triangle_i$:

$$y_i^\star = \arg\max_{y_i} P(y_i | Y^L, \mathbf{X}, G)$$

Experiment Setting

- Code&Data: http://arnetminer.org/decodetriad
- Data Set
  - Coauthor network from ArnetMiner[1]
  - Year span: 1995 - 2014
  - Formation time: the earliest year that two authors collaborate
  - 631,463 closed triads, 200,891 nodes
- Local Features
  - Demographic features: #pubs and #collaborators for each author
  - Interaction features: #common-pubs, #common-conferences, etc. for each pair of authors
  - Social effect features: PageRank score and structural hole spanner score[2] of each author

[1] https://aminer.org/
[2] Lou, T., & Tang, J. Mining structural hole spanners through information diffusion in social networks. WWW'13.

Decoding Performance

>20% improvement in terms of accuracy

Algorithm	Spearman	Kendall	Accuracy
Rule	0.4604	0.3525	0.3293
SVM	0.3205	0.2286	0.4121
Logistic	0.3379	0.2407	0.4830
DeTriad-A	0.3060	0.2190	0.5550
DeTriad	**0.2716**	**0.1935**	**0.5964**

Rule: Rank edges directly by the number of coauthor papers on each edge.

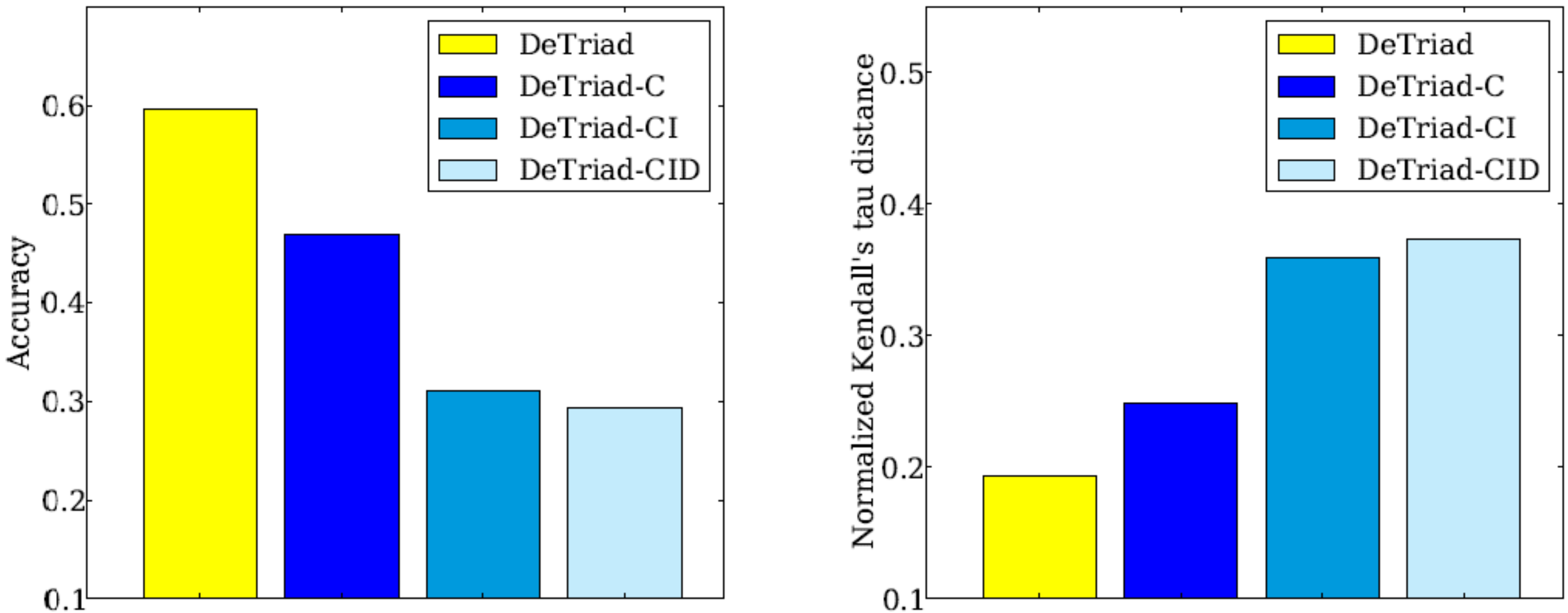
SVM: Support Vetor Machine using local features.

Logistic: Logistic Regression using local features.

DeTriad-A: **DeTriad** defined by an asynchronous method.

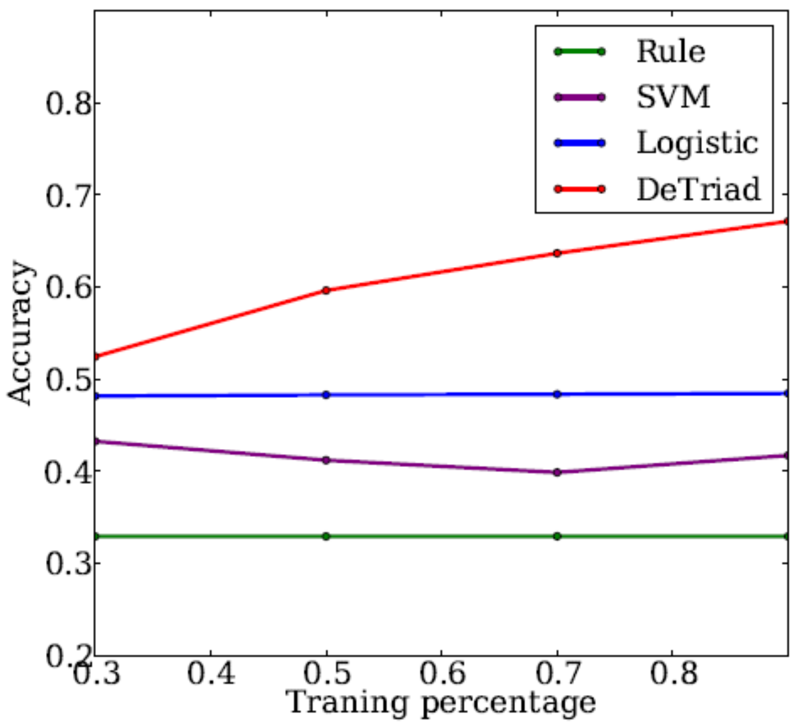
DeTriad: **DeTriad** defined by a synchronous method.

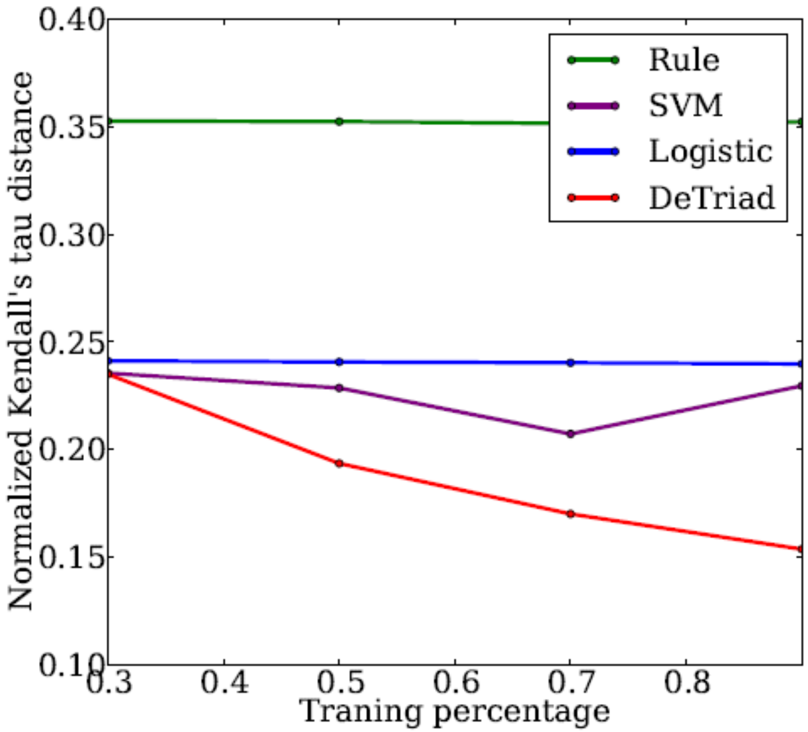
Factor Contribution Analysis



DeTriad-C: stands for removing correlation features
DeTriad-CI: stands for further removing interaction features
DeTriad-CID: stands for further removing demography features

Performance with Different Train/Test Ratio

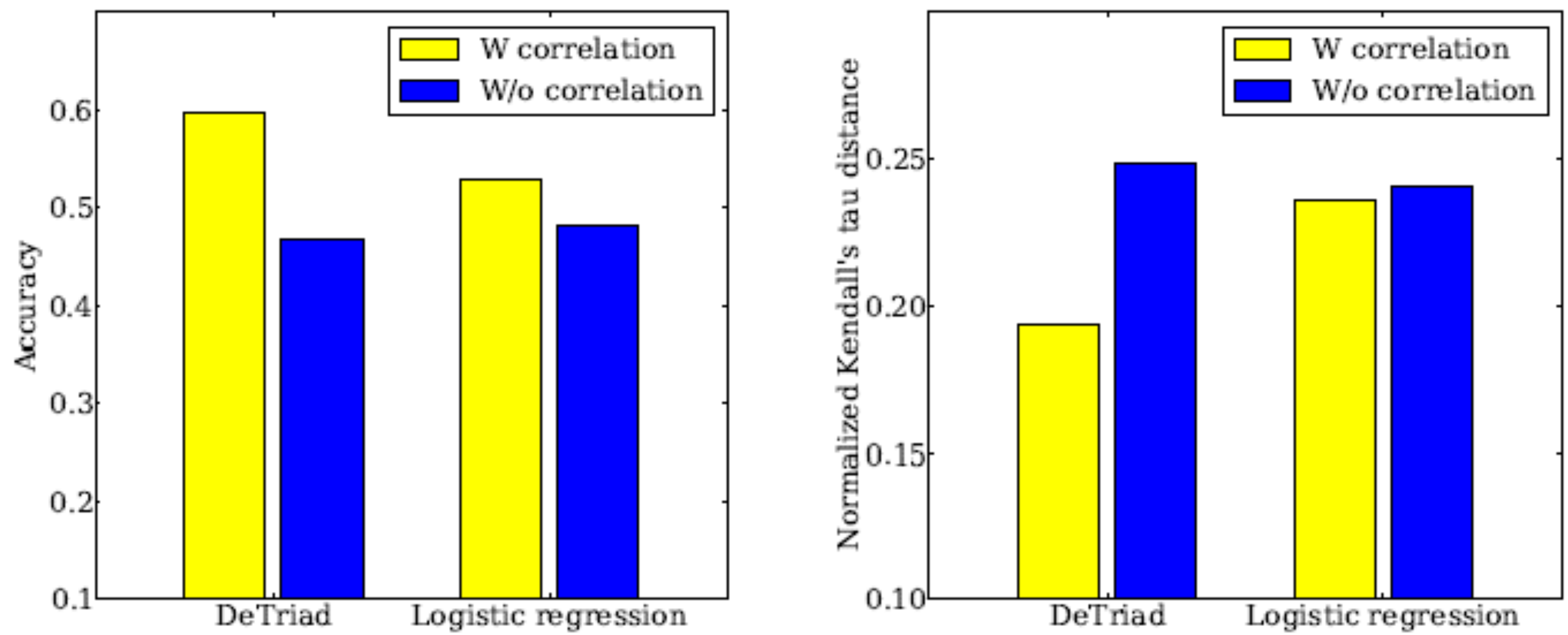




DeTriad can capture more information from large training data because of the correlation factors

Effect of Correlation Factors

- Compare to LRC with correlation features
  - Use the # of labeled triads that an edge is the k^{th} formed edge for LRC correlation features



Correlation factors better model the correlation among triads

Conclusion

- Formulate the problem of decoding triadic closures.
- Propose the DeTriad model integrating correlations among closed triads and partial labeled information to solve this problem.
- Show that our model outperforms several alternative methods by up to 20% in terms of accuracy.

Thanks!

Jie Tang, KEG, Tsinghua U, http://keg.cs.tsinghua.edu.cn/jietang
Download data & Codes, http://arnetminer.org/decodetriad