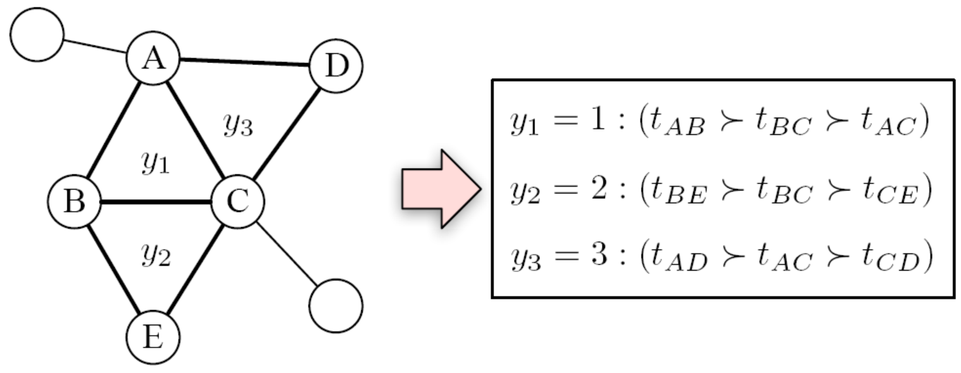
Uncovering the Formation of Triadic Closure in Social Networks

Zhanpeng Fang and Jie Tang
Department of Computer Science, Tsinghua University

- Group formation is a central research issue in social science. A triad is a group of three people, one of the most basic human groups in social networks.
- Understanding the formation of triads can help reveal the complex and subtle mechanism that governs dynamics of all social networks.
- Decoding triadic closure: for a given closed triad, predict which link was firstly created, which followed and finally which one was the last to close the triad.

- **DeTriad**: a semi-supervised probabilistic factor graph model which naturally models the correlation between triads, outperforms several alternative methods by 20% in terms of accuracy.



Problem Definition

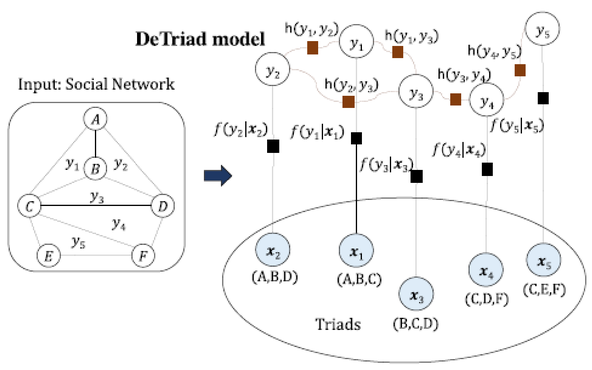
Let $G = (V, E)$ denote a social network. Associate a time stamp $t_{i,j}$ to each edge.

Definition 1 ***Closed Triad:*** *For three users $\Delta = (A, B, C)$, if there is a link between any two users, i.e., $e_{AB}, e_{BC}, e_{AC} \in E$, then we say that Δ is a closed triad.*

Problem 1 ***Decoding Triadic Closure:***
Given a social network $G = (V, E)$, associate a hidden variable y_i to each closed triad Δ_i to represent the order of time stamps associated with the three links in Δ_i.
Goal: use available labeled information to train a function f, so that, for an unlabeled Δ, we can predict its order y, i.e.,

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

DeTriad Model



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

- **Local factor:** $f(y_i|\boldsymbol{x}_i)$ represents posterior probability of the decoding result given all features defined for triad Δ_i;

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

- **Correlation factor:** $h(y_i, y_j)$ represents the correlation between decoding results of Δ_i and Δ_j.

$$h(y_i, y_j) = \frac{1}{Z_2} \exp\{\sum_k \mu_k h_k(y_i, y_j)\}$$

- Synchronous method for correlation factor

$$h_(y_i, y_j) = \frac{1}{Z_3} \exp\{\sum_k \mu_k \cdot I_k(y_i, y_j)\},$$

- Asynchronous method for correlation factor

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

$$\mathcal{O}(\boldsymbol{\theta}) = \log P(Y^L|\boldsymbol{X}, G) = \log \sum_{Y|Y^L} P(Y|\boldsymbol{X}, G)$$

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

$$- \log \sum_{Y} \{\sum_{\Delta_i} \sum_{k=1}^{d} \alpha_k f_k(x_{ik}, y_i) + \sum_{i \sim j} \sum_k \mu_k h_k(y_i, y_j)\}$$

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

Input: network G, features $\boldsymbol{X}$, learning rate η
Output: estimated parameters $\boldsymbol{\theta}$

Initialize $\boldsymbol{\theta} \leftarrow \mathbf{0}$;
repeat
  Perform LBP to calculate $P(y_i|Y^L, \boldsymbol{X}, G)$, $P(y_i, y_j|Y^L, \boldsymbol{X}, G)$;
  Perform LBP to calculate $P(y_i|\boldsymbol{X}, G)$, $P(y_i, y_j|\boldsymbol{X}, G)$;
  Calculate the gradient ∇_{μ_k} of μ_k according to Eq. 8 (for α_k with a similar formula):

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

  Update parameter $\boldsymbol{\theta}$ with the learning rate η:

$$\boldsymbol{\theta}_{\text{new}} = \boldsymbol{\theta}_{\text{old}} + \eta \cdot \nabla_{\boldsymbol{\theta}}$$

until *Convergence*;

Algorithm 1: Learning algorithm for the DeTriad model.

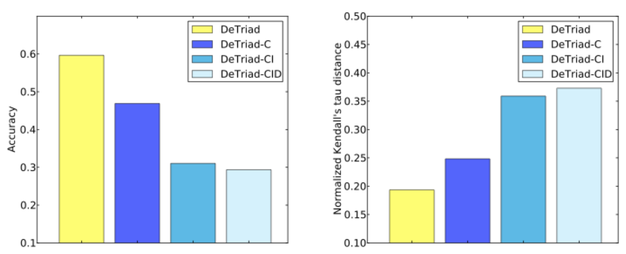
Experimental Results

Dataset: coauthor network from 1995 to 2014 in Arnetminer includes 1,910,979 publications, 1,145,632 authors, 4,322,998 coauthor relationships, and 5,913,455 closed triads.

4 Categories of features: C(Correlation), D(Demography), I(Interaction), S(Social effect)

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**

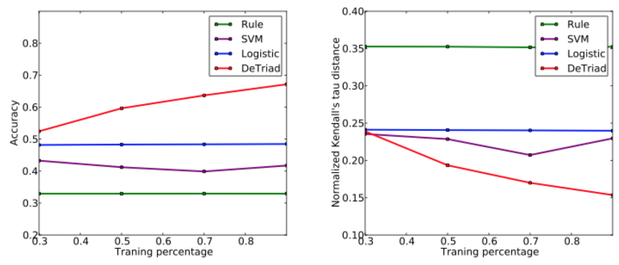
Factor Contribution:



Each category of features contributes improvement to the performance.

Effect of Correlation Factors:
The improvement of DeTriad by adding correlation factors is significantly higher than that of logistic regression by adding correlation features.

Training/Test Ration:



Improvements can be obtained by DeTriad when increasing the size of training data.