# Fuzzy models for link prediction in social network

**Ahmad Khalili Jafarabad**

International Summer School on Social Network Analysis(2013)
National Research University Higher School of Economics

# Link pridiction

- The relevant problems of link prediction are more concerned to the links that will be formed in the future or finding the missing links in the network.
- The link prediction based upon the graphs' structure does not take real world factors such as users' characteristics and the root of relations into account (Backstrom and Leskovec 2011)

# Application

- Predicting the Structure and missing links of the covert networks or terrorists(Basu 2005)
- In biomedical problems which measuring the node atributes are not easy.
- Working with vague or noisy data.

# Why Fuzzy modeling

- same age, geographical place, more common friends
- Most of the concepts which are using in SNA are liguistic( close, weak, most,strong,…)

# Fuzzy VS. Crisp

- Crisp Relations

$$R(x_1, x_2, \ldots x_n) = \begin{cases} 1, & \text{iff } < x_1, x_2, \ldots x_n > \in R \\ 0, & \text{otherwise} \end{cases}$$

- Fuzzy Relations

$$\mu_R(x_i, x_j) = \begin{cases} 1 & \text{if } x_i \textit{ has the strongest possible degree of relationship with } x_j \\ \gamma \epsilon ]0,1[ & \text{if } x_i \text{ to a certain extent, related to } x_j \\ 0 & \text{if } x_i \text{ is not related with } x_j \end{cases}$$

# Literature Review

| Name and formulation | | Reference |
|---|---|---|
| Local similarity index | | |
| Common neighbors | $\lvert\Gamma(x) \cap \Gamma(y)\rvert$ | **(Newman 2001)** |
| Jaccard | $\frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$ | **(Jaccard 1901)** |
| Adamic-Adar | $\frac{1}{log(\lvert\Gamma(z)\rvert)}$ $z \in \Gamma(x) \cap \Gamma(y)$ | **(Adamic and Adar 2003)** |
| Sørensen | $\frac{2\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x)\rvert + \lvert\Gamma(y)\rvert}$ | **(Sørensen 1948)** |
| Prefrential attachements | $\lvert\Gamma(x)\rvert . \lvert\Gamma(y)\rvert$ | **(Barabási and Albert 1999)** |
| Salton | $\frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\sqrt{k_x k_y}}$ | **(Salton and McGill 1986)** |
| Cluttering Coefficient | $\frac{2t_x}{\lvert\Gamma(x)\rvert * \lvert\Gamma(x)-1\rvert} + \frac{2t_y}{\lvert\Gamma(y)\rvert * \lvert\Gamma(y)-1\rvert}$ | **(Watts and Strogatz 1998)** |
| Hub promoted Index | $\frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{min\{k_x, k_y\}}$ | **(Ravasz, Somera et al. 2002)** |

# Literrature Review

| | | |
|---|---|---|
| Resource alllocation | $\frac{1}{k_x}$ | **(Zhou, Lü et al. 2009)** |
| Global similarity index | | |
| Katz | $\sum_{l=1}^{\infty} \beta^l . \lvert paths_{xy}^{<l>} \rvert = \beta A_{xy} + \beta^2 (A_{xy}^2) + \beta^3 (A_{xy}^3) + \cdots$ | **(Katz 1953)** |
| Average commute time | $M(x,y)+M(y,x)$ | **(Fouss, Pirotte et al. 2005)** |
| Leicht-Holme-Newman index | $\varphi(I + \varphi A + \varphi^2 A^2 + \cdots)$ | **(Leicht, Holme et al. 2006)** |
| Quasi local index | | |
| Local path | $A^2 + \varepsilon A^3$ | **(Zhou, Lü et al. 2009)** |

# Checking Accuracy

- **AUC (Area Under the receiver operating characteristic curve ):**
- generates a score for all of the non-existing links in each step and makes a relationship between two nodes with the highest score.
- Some of the edges are removed to create the  set that is the set of <u>non-observed links</u>.
- Therefore, the score of these links are compared with the score of <u>non-existing links</u>.
- In *n* independent comparison, *n′* counts the times that score of non-observed links is higher and *n″* counts vice versa, then *AUC* will be computed as follows:

$$AUC = \frac{\acute{n} + \mathrm{n}''}{\mathrm{n}}$$

# Fuzzy clustering coefficient

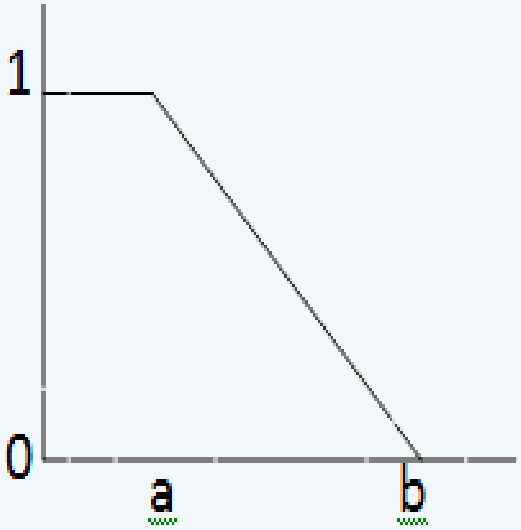
The clustering coefficient is based on the clique concept

**softer definitions for clustering coefficient (Yager 2008):**

- "Most of the elements in S are closely connected".
- ""None of the elements in S are too far from any of the others".
- ""No element not on the clique is better connected to the members of clique than any element in the clique".

# Fuzzy clustering coefficient

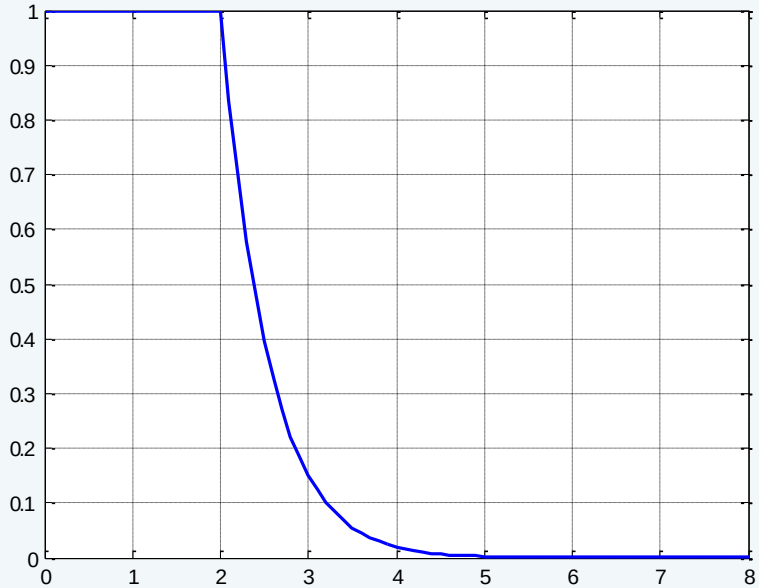
1. Close:
   1. It can be defined as a path with minimum length that connects two nodes to each other.
   2. Yager conceptual model for close:



# Fuzzy clustering coefficient

- New close function
  - the closeness in social networks decreases exponentially.
  - new close function for undirected and unweighted social networks is proposed as follows:

$$close(i,j) = \frac{q_{ij}}{2 * 10^{q_{ij}-2}}$$



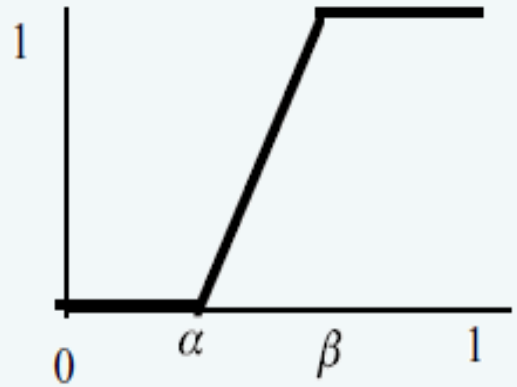
# Fuzzy clustering coefficient

Closeness for fuzzy graphs:

$$\begin{cases} 1 & q_{ij} < 2 \\ \dfrac{(w(i,z) + w(z,j))}{2 * 10^{q_{ij}-2}} & q_{ij} = 2 \\ \dfrac{\big(w(i,z) + w(z,e) + w(e,j)\big)}{2 * 10^{q_{ij}-2}} & q_{ij} = 3 \\ 0 & q_{ij} > 3 \end{cases}$$

# Fuzzy clustering coefficient

- Most:
- Most can also be defined as a fuzzy function like *M(p)* that indicates the proportion *p* satisfies the Most (Yager 2008).




$$\begin{cases} M(p) = 0 & p \leq \alpha \\ M(p) = \dfrac{\beta - p}{\beta - \alpha} & \alpha \leq p \leq \beta \\ M(p) = 1 & p \geq \beta \end{cases}$$

- For any node in the network the value of *p* can be computed as follows:

$$p_{x_i} = \frac{\sum_{\substack{j=1 \text{ to } n_s \\ i \neq j}} close(x_i, x_j)}{n_s - 1}$$

- C1 criteria can be calculated by using the amount of

$$M(p_{x_i})$$

- there are *Far* and *Not Far* concepts.
- *Far* is a function like the close and the *Not* indicates the negation of this fuzzy number. For any pairs of nodes *Not Far* is calculated by equation:

$$Not.Far(x,y) = Max_{k=1\ to\ n}[R^k(x,y) \wedge \bar{F}(k)]$$

- Second Criteria:

$$C_2(S) = MIN_{u \epsilon U_S}[Not.Far(u)]$$

The third criterion shows that every node out of the cluster of considered nodes should not be closer to most of the nodes in the cluster. This criterion should be defined as follows (Yager 2008)

$$M(y/S) = Most\left(\frac{\sum_{j=1}^{n_s} close(y, x_j)}{n_s}\right)$$

$$M(x_i/S) = Most\left(\frac{\sum_{\substack{j=1\ to\ n_s \\ i \neq j}} close(x_i, x_j)}{n_s - 1}\right)$$

In the equations $x_i \in S$ و $y \notin S$

If M(y/s) is lower than the M(x/s) for all of the nodes in the cluster then the third criteria is 0 otherwise 1.

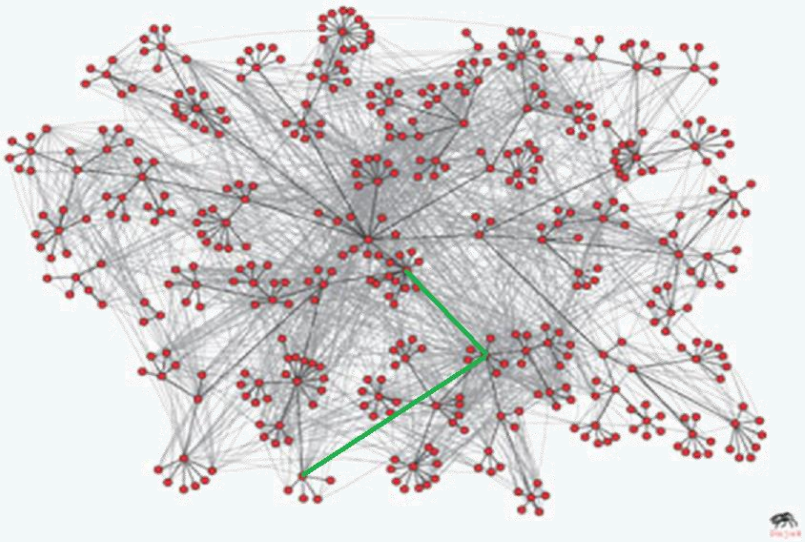
## Fuzzy link prediction based on local clustering coefficient (FCC)

- fuzzy quasi-local clustering coefficient model is used for link prediction.
- The score of the node in the paper is calculated similar to the previous studies in the literature (Saramäki, Kivelä et al. 2007)

$$S(x,y)=S(x)+S(y)$$

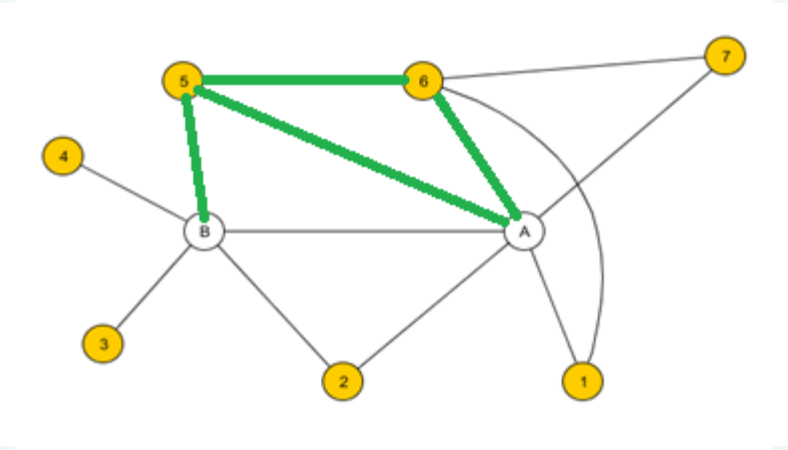
# Fuzzy link prediction based on cluster overlapping(FCO)

- In previous models based on clustering coefficients or cluster of the node, the clusters' overlapping is not taken into account



# Fuzzy link prediction based on cluster overlapping(FCO)

- Goldberg proposes a model to find the cluster overlaps based on counting the nodes that are similar in both clusters over all of the nodes of both clusters (Goldberg, Hayvanovych et al. 2010).

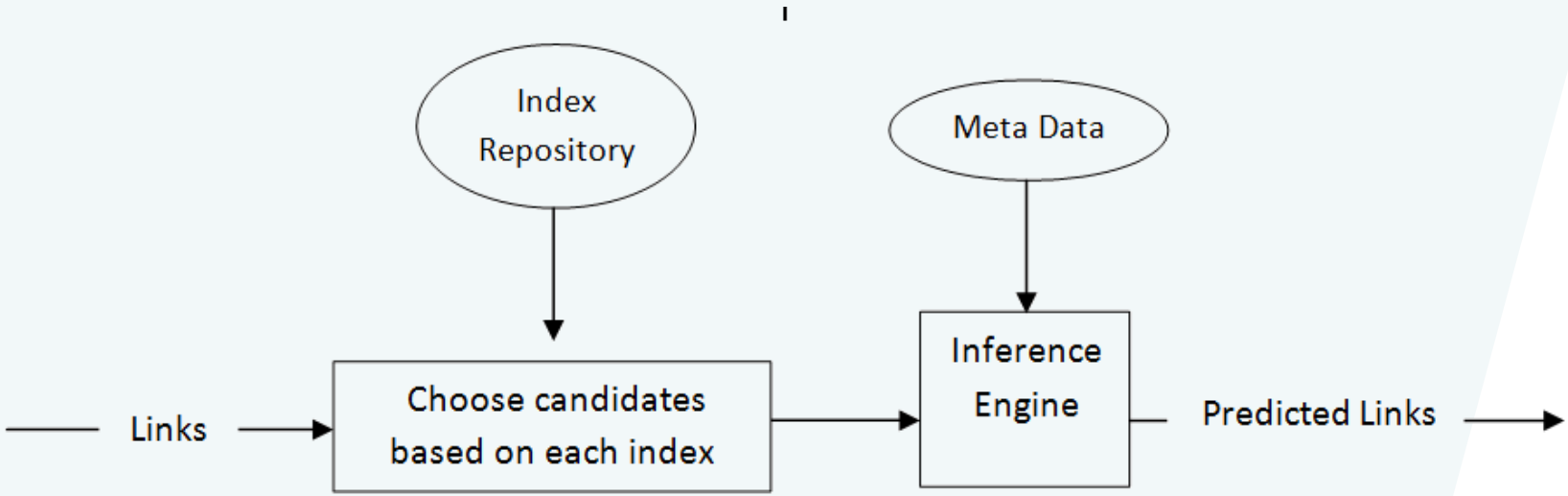


- In order to compute the cluster overlapping for two nodes and the following new index is proposed:

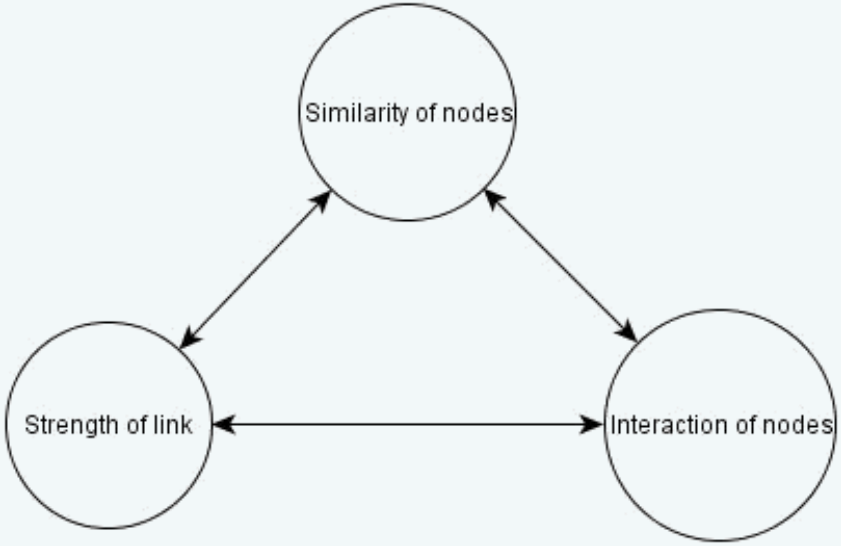
$$\frac{\sum_{z=1}^{n} close\left(x_i, x_j\right)}{\left|\sum_{u,z \in S_{x_i}} W_{uz}\right| + \left|\sum_{u,z \in S_{x_j}} W_{uz}\right|}.$$

- Equation calculates the sum of closeness of two nodes through all the paths that connect the nodes to each other over the sum of weights of all edges within the clusters

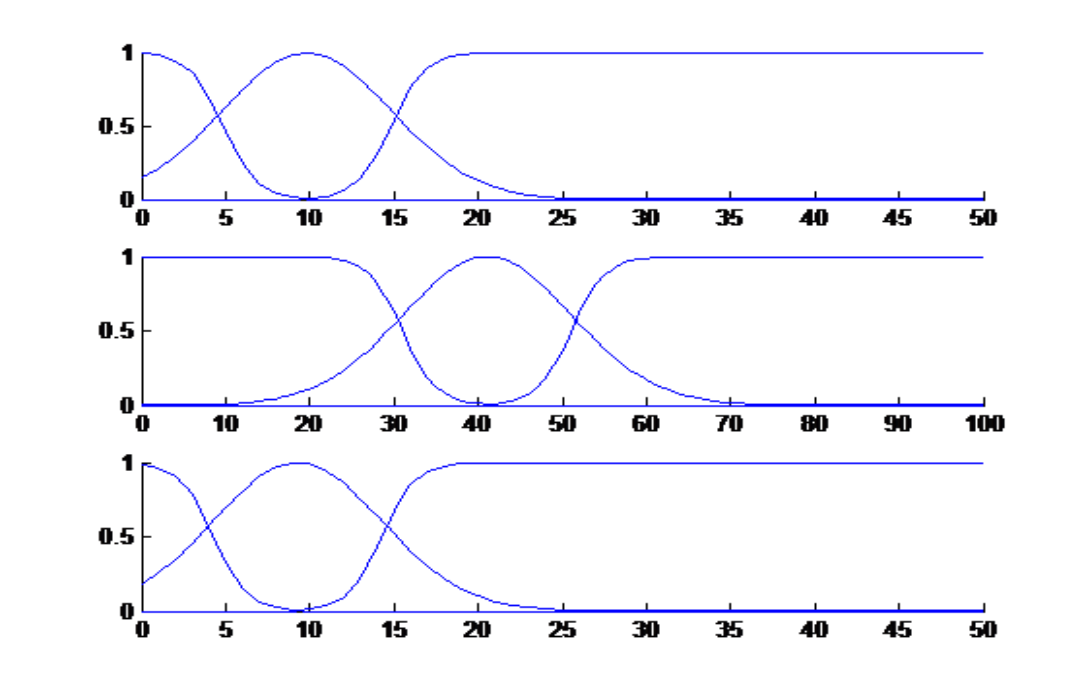
# Intelligent hybrid methodology for link prediction

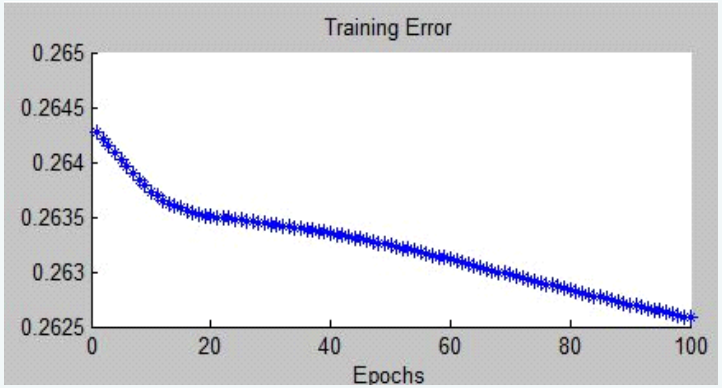


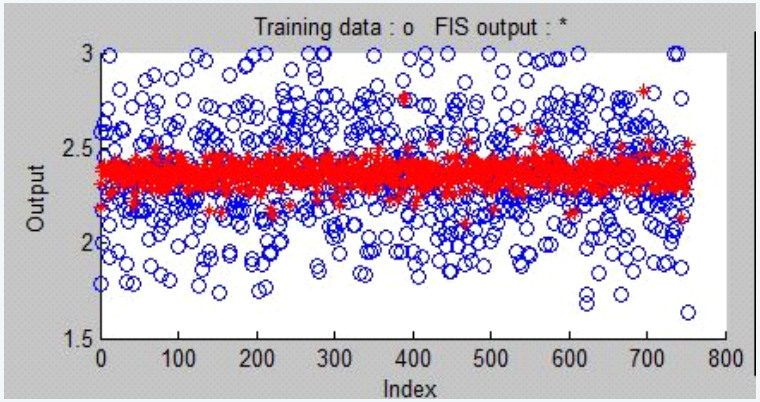
- The similarity indicates the characteristics of people like age, gender, location
- Interaction shows the frequency of the relations between two nodes.



# MAMDANI-ANFIS inference engine




Training Error
0.265
0.2645
0.264
0.2635
0.263
0.2625
0
20
40
60
80
100
Epochs


Training data : o FIS output : *
3
2.5
2
1.5
Output
0
100
200
300
400
500
600
700
800
Index

# Choosing the best candidate

- different proximity measures may choose the same candidates for creating the links
- the frequency of the selection and the weights of candidates is taken into consideration.
- In the equation w is the strength of a link, f the number of selections and N the number of all indices.

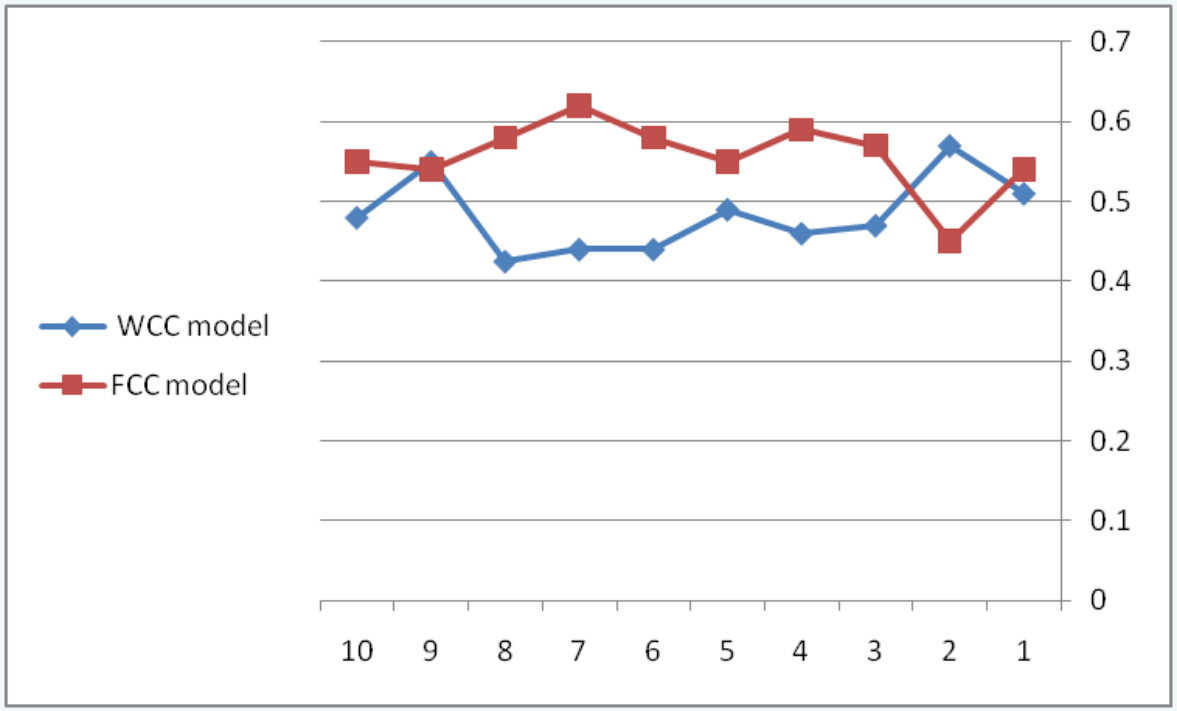
$$S_F = max_{for\ all\ candidates} \left(w_{ij} * \frac{f_i}{N}\right)$$

# DATA

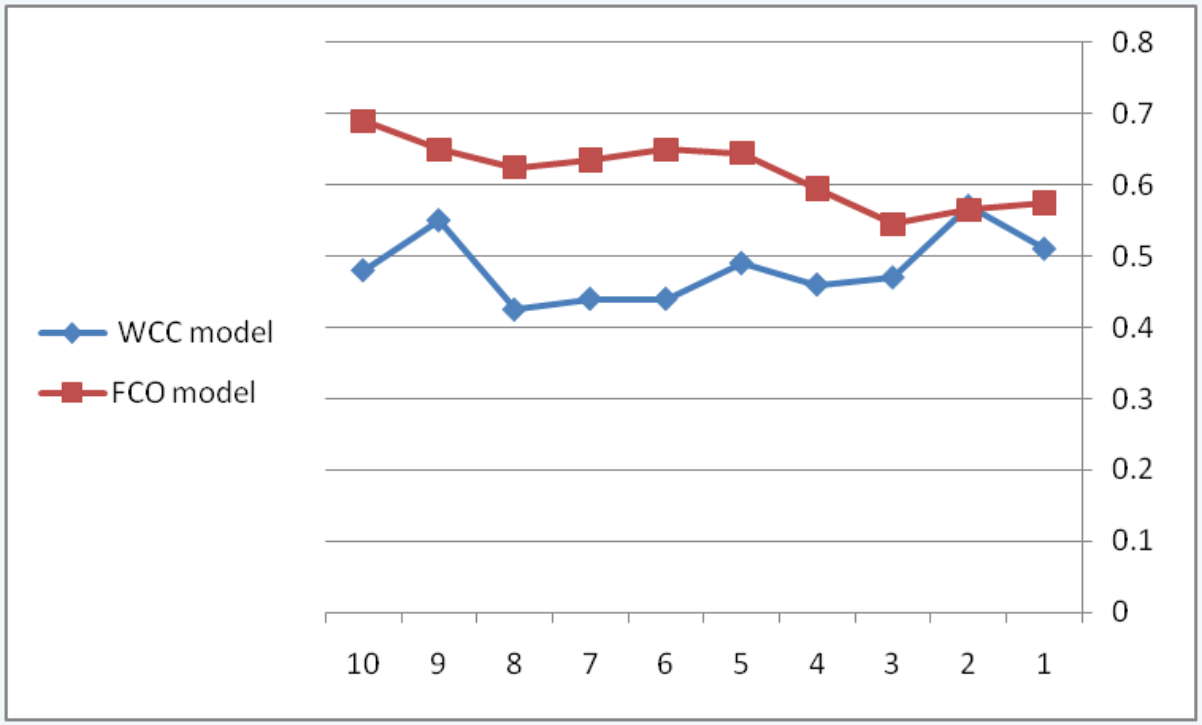
- Netscience coauthor ship network
- Coauthor ship network data set(Newman 2006)
- The weights are calculated by the number of works which is done by two researchers in common.
- In every step about 10 percent of the data were chosen for the test
- for every model about 10 experiments were done

# Resutls

- FCC VS. WCC:

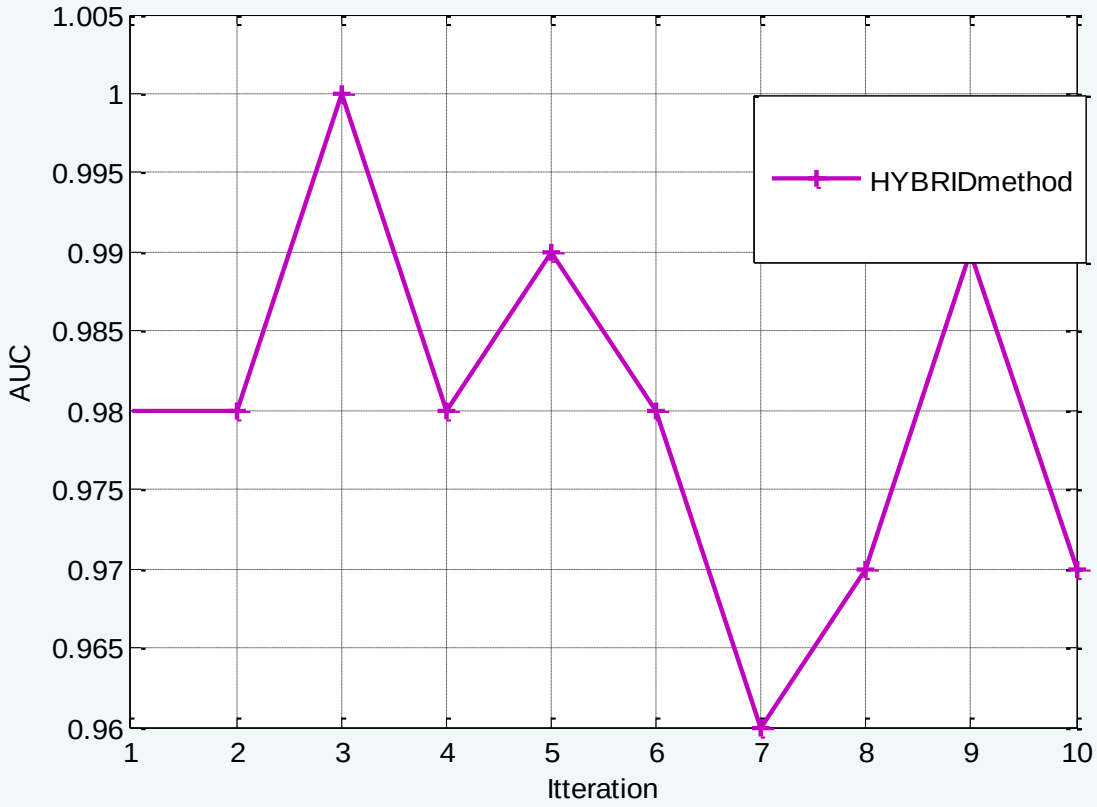


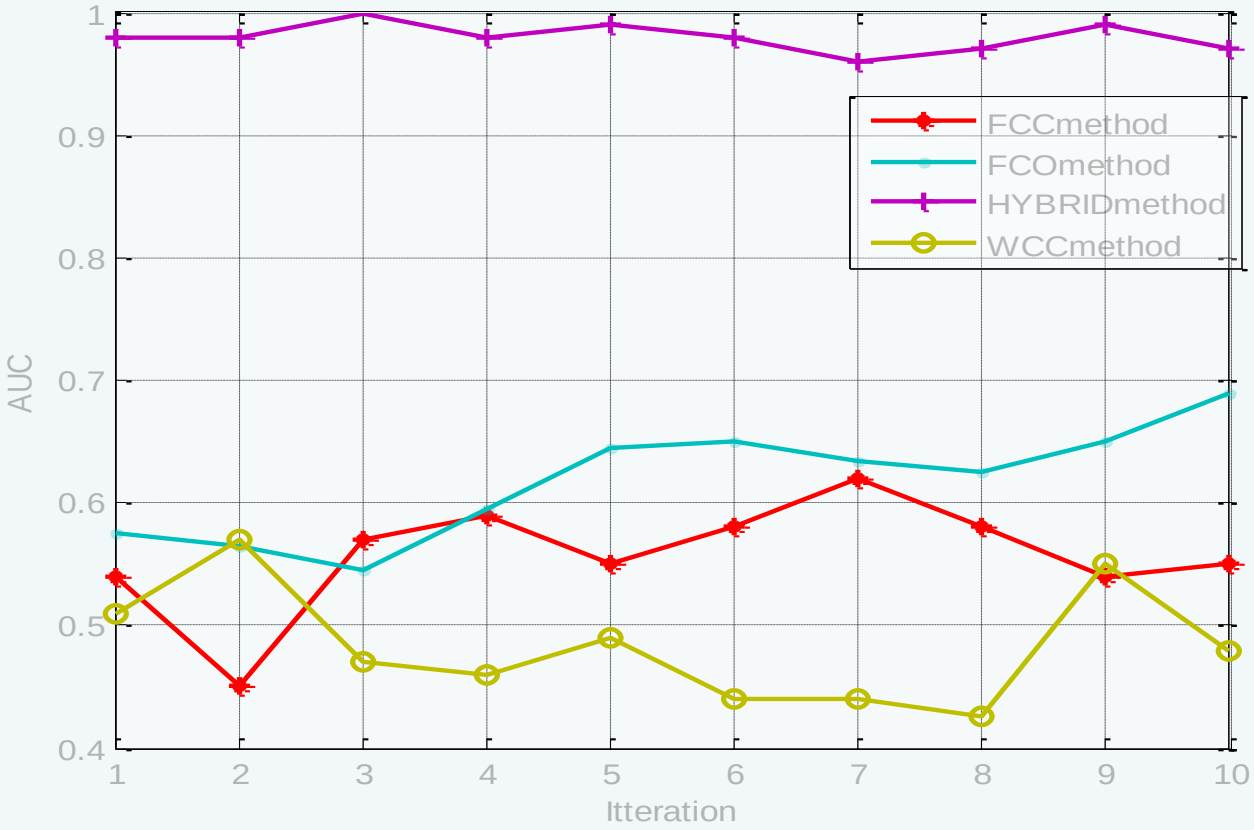
## ❖ FCO VS. WCC:



# Results

- Hybrid Model




FCCmethod
FCOmethod
HYBRIDmethod
WCCmethod
AUC
Itteration
1
0.9
0.8
0.7
0.6
0.5
0.4
1
2
3
4
5
6
7
8
9
10

# Conclusion and future works

- The results show that using fuzzy logic can make better predictions because of the better definition of networks characteristics and the concepts related to the links and nodes in social networks.
- It also shows that the hybrid model presents the best predictions, and cluster overlapping models has the second best accuracy in predictions.
- The results also indicate that ANFIS is weak in predicting the strength of the links.
- It also seems that Mamdani inference engine cannot predict the strength of nodes due to the complexity of the social networks

# Conclusion and future works

- The current study is the first attempt in the domain of fuzzy link prediction.
- there are rooms of works that can be done in future
- it is possible to use some models based on machine learning techniques such as fuzzy logistic regression and fuzzy EM (expectation maximization) to predict the strength of the links more effectively.
- Weak ties theory, which plays an important role in social networks, and fuzzy logic can model the strength of the ties very well.
- It is worth noting that employing the fuzzy probability based models can modify the methods for finding the strength of the links or predicting the evolutions of the social networks.

# Publication:

- Bastani, Susan, Ahmad Khalili Jafarabad, and Mohammad Hossein Fazel Zarandi. "Fuzzy Models for Link Prediction in Social Networks." *International Journal of Intelligent Systems* (2013).

# Thanks!