Exploiting behaviors of communities of twitter users for Link prediction

[Authors: Jorge Valverde-Rebaza, Alneu de Andrade Lopes]

Raaghav Radhakrishnan
246097

Data Analytics Seminar I

18.12.2018

Overview

- Introduction
- Problem description
  - Unsupervised strategy
  - Supervised strategy
- Structural link prediction measures
- Prediction with community information
  - Label Propagation algorithm
  - WIC measure
  - W form measures
- Experiments and Results
  - Experimental setup
  - Results and analysis
- Conclusion
- Recommendation
- References

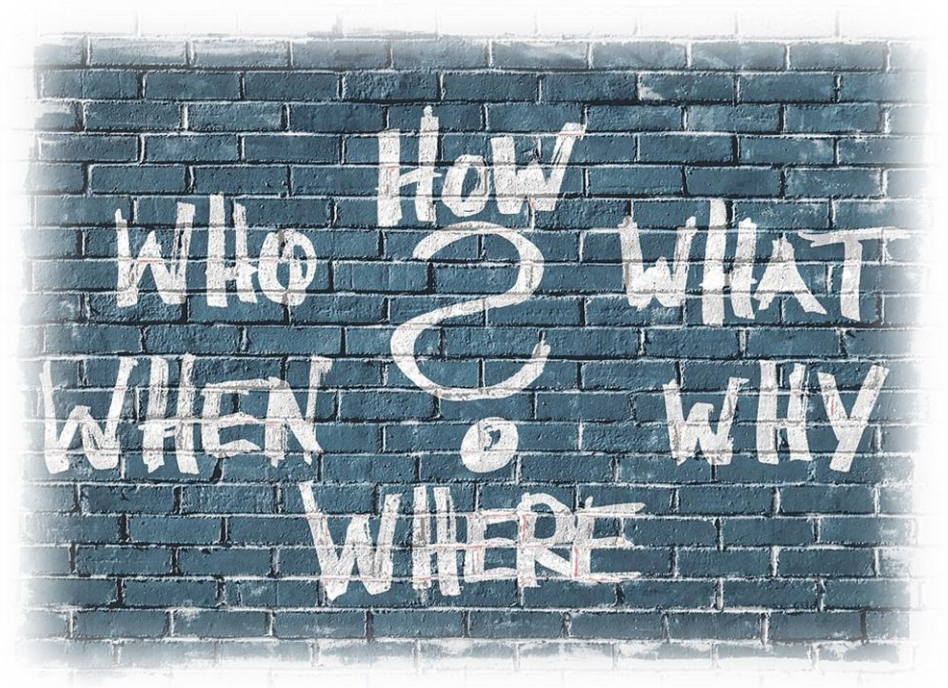
Overview

- Introduction
- Problem description
  - Unsupervised strategy
  - Supervised strategy
- Structural link prediction measures
- Prediction with community information
  - Label Propagation algorithm
  - WIC measure
  - W form measures
- Experiments and Results
  - Experimental setup
  - Results and analysis
- Conclusion
- Recommendation
- References

Introduction

Link Prediction

- What
- Who
- Why
- Where
- How



Introduction

- New interactions among members
- Tool to find hidden social relationship

- Users of social networking services
- Users of Facebook and Twitter

- To allow individuals network with others with similar personal and business interests

- Bioinformatics
- E-commerce
- Security
- Collaborative

- Link Prediction problem:
  - Supervised
  - Unsupervised

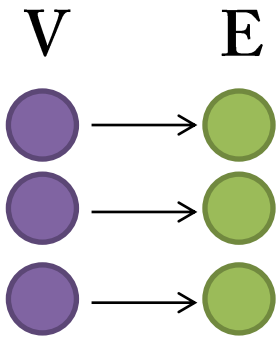
Overview

- Introduction
- Problem description
  - Unsupervised strategy
  - Supervised strategy
- Structural link prediction measures
- Prediction with community information
  - Label Propagation algorithm
  - WIC measure
  - W form measures
- Experiments and Results
  - Experimental setup
  - Results and analysis
- Conclusion
- Recommendation
- References

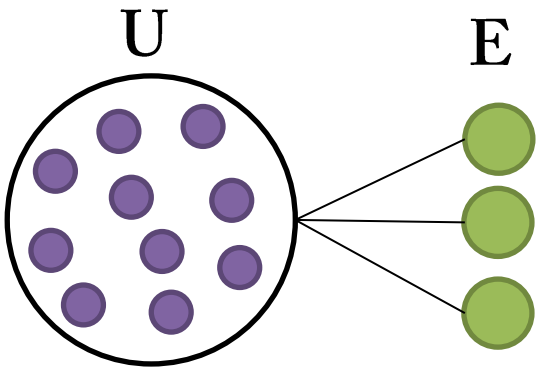
Problem description

- Previous techniques for link prediction were based on structural (or topological) information
- Structural information is not enough to achieve a good performance in the link prediction task on large-scale social networks.
- The use of additional information, such as interests or behaviours that nodes have into their communities, may improve the link prediction performance
- Using a set of simple and non-expensive techniques that combine structural with community information for predicting the existence of future links in a large-scale online social network

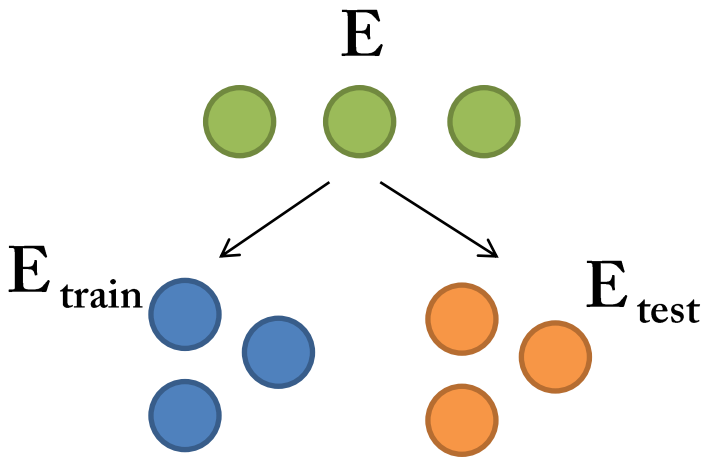
Problem description



V ⟶ Set of nodes
E ⟶ Links of nodes in V



U ⟶ Set of all nodes



Problem description

Prediction accuracy for unsupervised strategy:

- AUC (Area Under ROC Curve)
- Precision

Area Under ROC Curve (AUC):

What is an ROC Curve?

An **ROC curve** (**receiver operating characteristic curve**)shows the accuracy of the prediction model according to the list of scores.

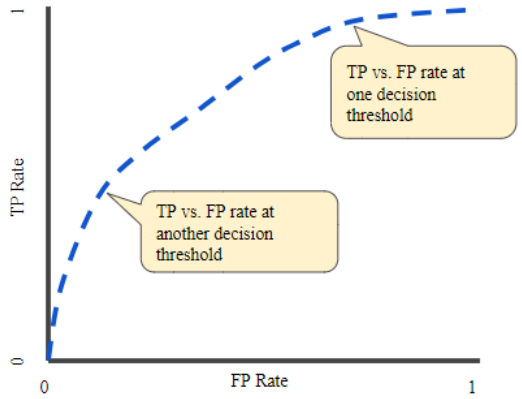
This curve plots: True Positive Rate & False Positive Rate

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

Problem description

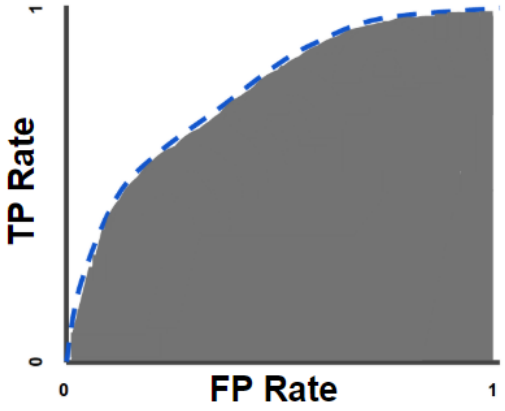
Area Under ROC Curve (AUC):

AUC measures the entire two-dimensional area underneath the entire ROC curve. AUC provides aggregate measure of the accuracy of prediction across the list of scores of all non-observed links.



ROC Curve

$$AUC = \frac{n' + 0.5n''}{n}$$



AUC

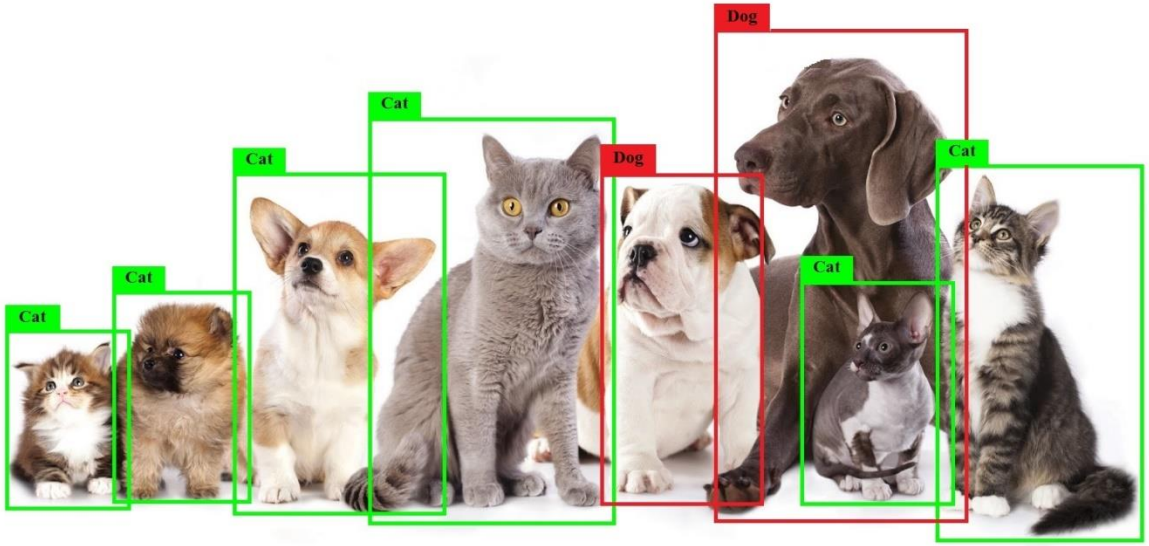
Precision:

Ratio between relevant items selected (L_T) and the items selected (L).

$$Precision = \frac{L_T}{L}$$

Problem description

Classifier performance for supervised strategy:



	Cat (True)		Dog (False)	
Actual	4		4	
Predicted	4 (TP)	2 (FP)	2 (TN)	0 (FN)

True Positive: Cat predicted as Cat
False Positive: Dog predicted as Cat
True Negative: Dog predicted as Dog
False Negative: Cat predicted as Dog

$$\text{Accuracy} = \frac{|\text{TP}| + |\text{TN}|}{|\text{TP}| + |\text{TN}| + |\text{FP}| + |\text{FN}|} \qquad \text{Precision}_S = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|} \qquad \text{Recall} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}$$

Overview

- Introduction
- Problem description
  - Unsupervised strategy
  - Supervised strategy
- Structural link prediction measures
- Prediction with community information
  - Label Propagation algorithm
  - WIC measure
  - W form measures
- Experiments and Results
  - Experimental setup
  - Results and analysis
- Conclusion
- Recommendation
- References

Structural link prediction measures

Structural measures use the similarity between nodes since similar nodes likely share same relations (links). Link prediction measures based on similarity can be classified into:

- ✓ Local structural information
- ✓ Global structural information

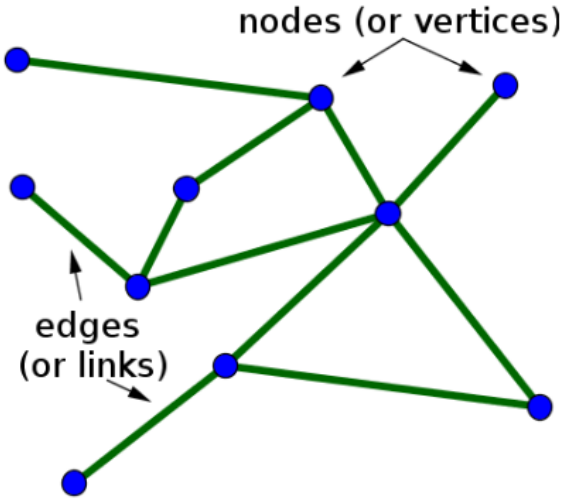
Global structural information	Local structural information
Pros: • Large amount of information • High accuracy	Pros: • Faster • Feasible for large-scale networks
Cons: • Time consuming • Infeasible for large-scale networks	Cons: • Only local information is available • Low accuracy

Structural link prediction measures

The basic structural definition for a node is its neighbourhood and it depends on the type of network formed by the link i.e., either directed or undirected.

Undirected network:

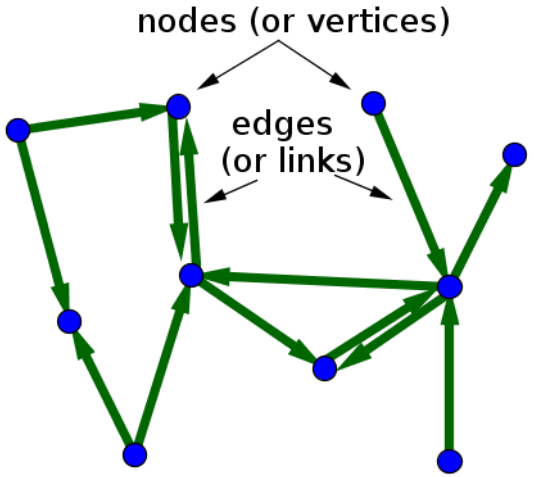
- Vertices or nodes are connected with bidirected edges.
- Example: Facebook and Linkedin, where once a request is accepted, both the nodes follow each other.



Structural link prediction measures

Directed network:

- Vertices or nodes are connected by directed edges
- Example: Twitter and Instagram, where once a request is accepted, it is not required that the other node should follow back
- Set of nodes formed by directed links from x (outgoing neighbourhood) is different from the set of nodes formed by directed links from them to x (incoming neighbourhood)



Structural link prediction measures

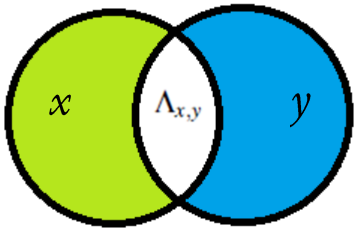
Local similarity measures:

1. Common neighbours (CN)
2. Adamic adar (AA)
3. Jaccard coefficient (Jac)
4. Resource allocation (RA)
5. Preferential attachment (PA)

$$s_{x,y}^{CN} = |\Lambda_{x,y}| = |\Gamma(x) \cap \Gamma(y)|$$

Common neighbours:

It refers to the size of set of all common friends of both x and y



Preferential attachment:

It is proportional to the number of neighbours of each vertex

$$s_{x,y}^{PA} = |\Gamma(x)| \times |\Gamma(y)|$$

Structural link prediction measures

Adamic Adar:

Refines sample counting of common neighbours by assigning one more weight to the less-connected neighbours.

$$s_{x,y}^{\mathrm{AA}} = \sum_{z \in \Lambda_{x,y}} \frac{1}{\log|\Gamma(z)|}$$

Jaccard Coefficient:

Indicates whether two users of a network have significant number of common neighbours regarding their total neighbours.

$$s_{x,y}^{\mathrm{Jac}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Resource Allocation:

Punishes high-degree common neighbours more heavily than AA

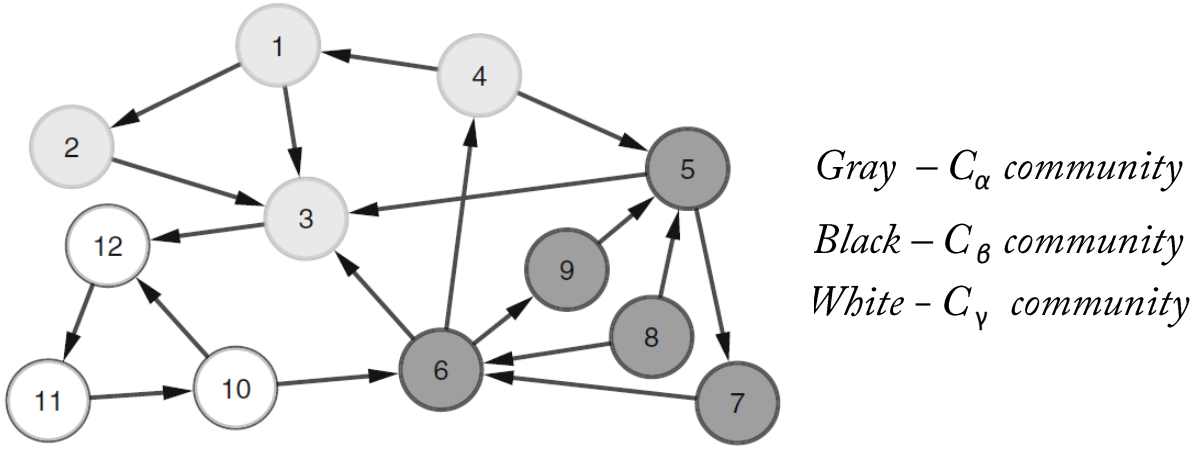
$$s_{x,y}^{\mathrm{RA}} = \sum_{z \in \Lambda_{x,y}} \frac{1}{|\Gamma(z)|}$$

Overview

- Introduction
- Problem description
  - Unsupervised strategy
  - Supervised strategy
- Structural link prediction measures
- Prediction with community information
  - Label Propagation algorithm
  - WIC measure
  - W form measures
- Experiments and Results
  - Experimental setup
  - Results and analysis
- Conclusion
- Recommendation
- References

Prediction with community information

- Consider in a network G, communities are represented by label $C_\alpha, C_\beta, \ldots, C_M$
- If $x \in V$ belongs to community with label C, then vertex is x^C



- Considering the above network with 12 vertices, the vertex 1 to 4 belongs to C_α, 5 to 9 belongs to C_β, 10 to 12 belongs to C_γ , the community detection process in large-scale social networks can be expensive. So, LPA is used for community detection

Prediction with community information

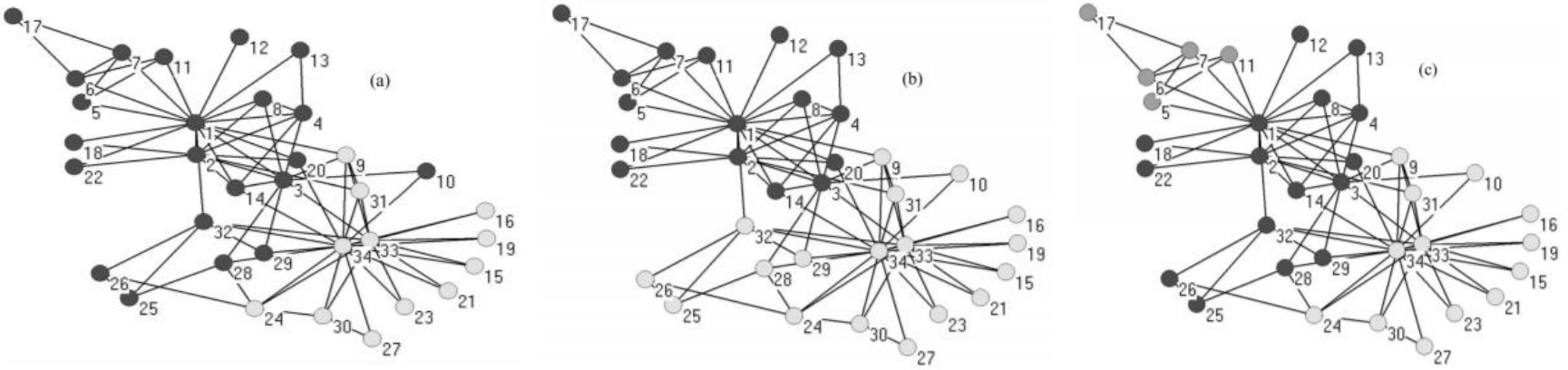
Label propagation algorithm:

Label propagation algorithm:

1: Initialize node label, $x, C_x = x$
2: Set $t=1$
3: Arrange nodes in random order
4: For $x \in X$, $C_x(t) = f(C_{xi1}(t), \ldots, C_{xim}(t), C_{xi(m+1)}(t-1), \ldots, C_{xik}(t-1))$
5: if
6: node label = max neighbour label
7: break
8: else
9: Set $t=t+1$

After this process, each node has more neighbours in its community than any other community

Prediction with community information



Zachary's karate club network which is a network of friendship among 34 members of a karate club 32 Over a period of time the club split into two factions due to leadership issues and each member joined one of the two factions. (a) – (c) are three different community structures identified by the algorithm on Zachary's karate club network. The communities can be identified by their shades of grey colours.

Advantages of LPA:

- Does not need information on the number and size of communities.
- Time complexity of algorithm is $O(m)$, where $m = |E|$
- Iterations to converge is independent of graph size

Prediction with community information

WIC Measure:

- Prediction measures based on structural similarity drastically improve as the community structure of network grows & vice versa
- WIC is used in large-scale networks
- Based on Bayesian theorem

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Deriving Bayesian theorem:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) \neq 0,$$

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}, \text{ if } P(A) \neq 0,$$

$$\Rightarrow P(A \cap B) = P(A \mid B) \times P(B) = P(B \mid A) \times P(A),$$

$$\Rightarrow P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}, \text{ if } P(B) \neq 0.$$

Prediction with community information

WIC Measure:

- Based on Bayesian theorem – given an undirected network G with M communities detected, the posterior probabilities that the same or different community labels are assigned to a pair of nodes (x,y)

$$P(x^{C_\alpha}, y^{C_\alpha} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) P(x^{C_\alpha}, y^{C_\alpha})}{P(\Lambda_{x,y})}$$

$$P(x^{C_\alpha}, y^{C_\beta} \mid \Lambda_{x,y}) = \frac{P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) P(x^{C_\alpha}, y^{C_\beta})}{P(\Lambda_{x,y})}$$

$$\Lambda_{x,y} = \Lambda^W_{x,y} \cup \Lambda^I_{x,y} \qquad \Lambda^I_{x,y} = \Lambda_{x,y} - \Lambda^W_{x,y}$$

$$\Lambda^W_{x,y} = \{z^C \in \Lambda_{x,y} \mid x^C, y^C\} \qquad \Lambda^W_{x,y} \cap \Lambda^I_{x,y} = \varnothing$$

- Probability of common neighbours $\Lambda_{x,y}$ given x^{Ca}, y^{Ca} and

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\alpha}) = \frac{|\Lambda^W_{x,y}|}{|\Lambda_{x,y}|}$$

- Probability of common neighbours $\Lambda_{x,y}$ given x^{Ca}, y^{Cb} are,

$$P(\Lambda_{x,y} \mid x^{C_\alpha}, y^{C_\beta}) = \frac{|\Lambda^{IC}_{x,y}|}{|\Lambda_{x,y}|}$$

Prediction with community information

Likelihood score of pair of vertices x and y is

$$s_{x,y} = \frac{\left|\Lambda^{W}_{x,y}\right|}{\left|\Lambda^{I}_{x,y}\right|} \times \frac{P(x^{C_\alpha}, y^{C_\alpha})}{P(x^{C_\alpha}, y^{C_\beta})} \qquad s^{WIC}_{x,y} = \frac{\left|\Lambda^{W}_{x,y}\right|}{\left|\Lambda^{IC}_{x,y}\right| + \delta}$$

In directed networks, the WIC measure of incoming and outgoing neighbourhood are,

$$s^{\mathrm{WIC_{in}}}_{x,y} = \begin{cases} \left|\Lambda^{W_{\mathrm{in}}}_{x,y}\right|, & \text{if } \Lambda^{W_{\mathrm{in}}}_{x,y} = \Lambda_{x,y} \\ \frac{\left|\Lambda^{W_{\mathrm{in}}}_{x,y}\right|}{\left|\Lambda^{I_{\mathrm{in}}}_{x,y}\right|}, & \text{otherwise} \end{cases} \qquad s^{\mathrm{WIC_{out}}}_{x,y} = \begin{cases} \left|\Lambda^{W_{\mathrm{out}}}_{x,y}\right|, & \text{if } \Lambda^{W_{\mathrm{out}}}_{x,y} = \Lambda_{x,y} \\ \frac{\left|\Lambda^{W_{\mathrm{out}}}_{x,y}\right|}{\left|\Lambda^{I_{\mathrm{out}}}_{x,y}\right|}, & \text{otherwise} \end{cases}$$

W form measures:

- Different common neighbours may give different contributions to the connection probability
- Within community common neighbours may contribute more to the connection likelihood than inter-community because they have similar behaviours. So, $\left|\Lambda^{W}_{x,y}\right|$ is used instead of $\Lambda_{x,y}$

Prediction with community information

Local measures	Undirected networks	Directed networks
Common neighbours	$s_{x,y}^{\mathrm{CN-W}} = \left\lvert \Lambda_{x,y}^{W} \right\rvert$	$s_{x,y}^{\mathrm{CN-W_{in}}} = \left\lvert \Lambda_{x,y}^{W_{\mathrm{in}}} \right\rvert$ $s_{x,y}^{\mathrm{CN-W_{out}}} = \left\lvert \Lambda_{x,y}^{W_{\mathrm{out}}} \right\rvert$
Adamic Adar	$s_{x,y}^{\mathrm{AA-W}} = \sum_{z \in \Lambda_{x,y}^{W}} \frac{1}{\log\lvert\Gamma(z)\rvert}$	$s_{x,y}^{\mathrm{AA-W_{in}}} = \sum_{z \in \Lambda_{x,y}^{W_{\mathrm{in}}}} \frac{1}{\log\lvert\Gamma_{\mathrm{in}}(z)\rvert}$ $s_{x,y}^{\mathrm{AA-W_{out}}} = \sum_{z \in \Lambda_{x,y}^{W_{\mathrm{out}}}} \frac{1}{\log\lvert\Gamma_{\mathrm{out}}(z)\rvert}$
Jaccard Coefficient	$s_{x,y}^{\mathrm{Jac-W}} = \frac{\left\lvert \Lambda_{x,y}^{W} \right\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$	$s_{x,y}^{\mathrm{Jac-W_{in}}} = \frac{\left\lvert \Lambda_{x,y}^{W_{\mathrm{in}}} \right\rvert}{\lvert\Gamma_{\mathrm{in}}(x) \cup \Gamma_{\mathrm{in}}(y)\rvert}$ $s_{x,y}^{\mathrm{Jac-W_{out}}} = \frac{\left\lvert \Lambda_{x,y}^{W_{\mathrm{out}}} \right\rvert}{\lvert\Gamma_{\mathrm{out}}(x) \cup \Gamma_{\mathrm{out}}(y)\rvert}$
Resource Allocation	$s_{x,y}^{\mathrm{RAW}} = \sum_{z \in \Lambda_{x,y}^{W}} \frac{1}{\lvert\Gamma(z)\rvert}$	$s_{x,y}^{\mathrm{RA-W_{in}}} = \sum_{z \in \Lambda_{x,y}^{W_{\mathrm{in}}}} \frac{1}{\lvert\Gamma_{\mathrm{in}}(z)\rvert}$ $s_{x,y}^{\mathrm{RA-W_{out}}} = \sum_{z \in \Lambda_{x,y}^{W_{\mathrm{out}}}} \frac{1}{\lvert\Gamma_{\mathrm{out}}(z)\rvert}$

Overview

- Introduction
- Problem description
  - Unsupervised strategy
  - Supervised strategy
- Structural link prediction measures
- Prediction with community information
  - Label Propagation algorithm
  - WIC measure
  - W form measures
- Experiments and Results
  - Experimental setup
  - Results and analysis
- Conclusion
- Recommendation
- References

Experiments and Results

Experimental Setup:

- Network Pre-processing
  - Set E is divided into training set and testing set
- Link Prediction process
  - Includes both unsupervised and supervised strategies

Results and analysis:

Table 1 Basic features of the two graphs built from Twitter network after 7th and 15th iterations of LPA algorithm

	Twitter 7it	Twitter 15it
$\|V\|$	24617334	24617333
$\|E\|$	363565896	363565892
M	3415051	2250964
Max cluster size	1392411	10121242
Ratio of total links per user	14.77	14.77

Experiments and Results

Execution time, in seconds, of all link prediction measure used in our experiments in the unsupervised strategy

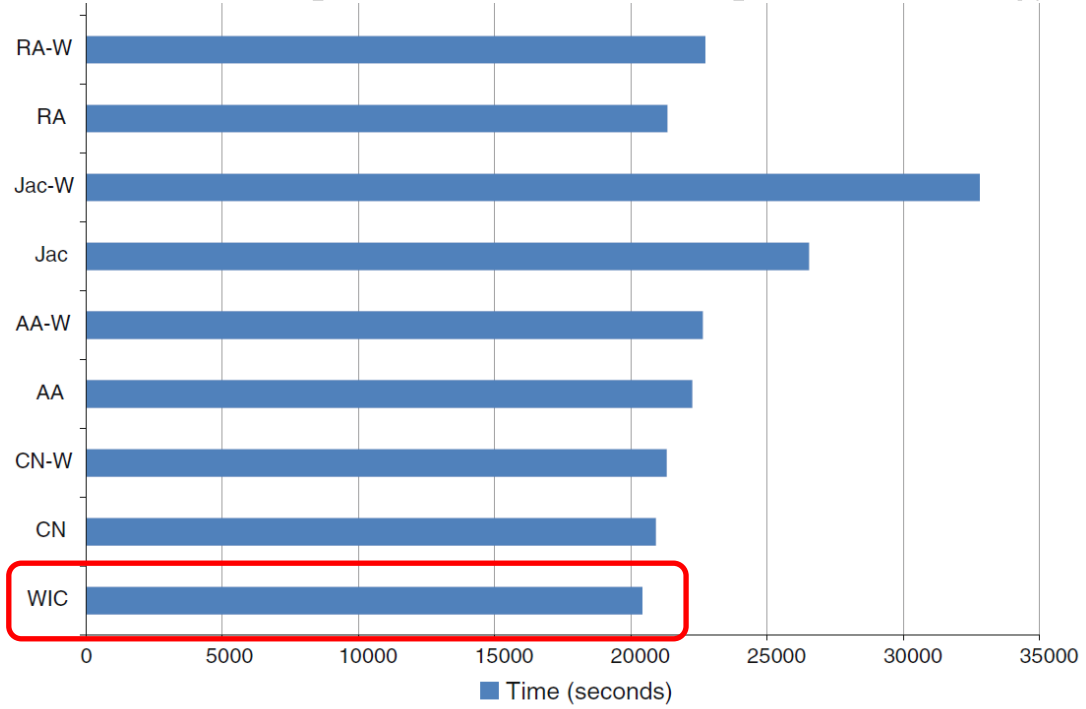
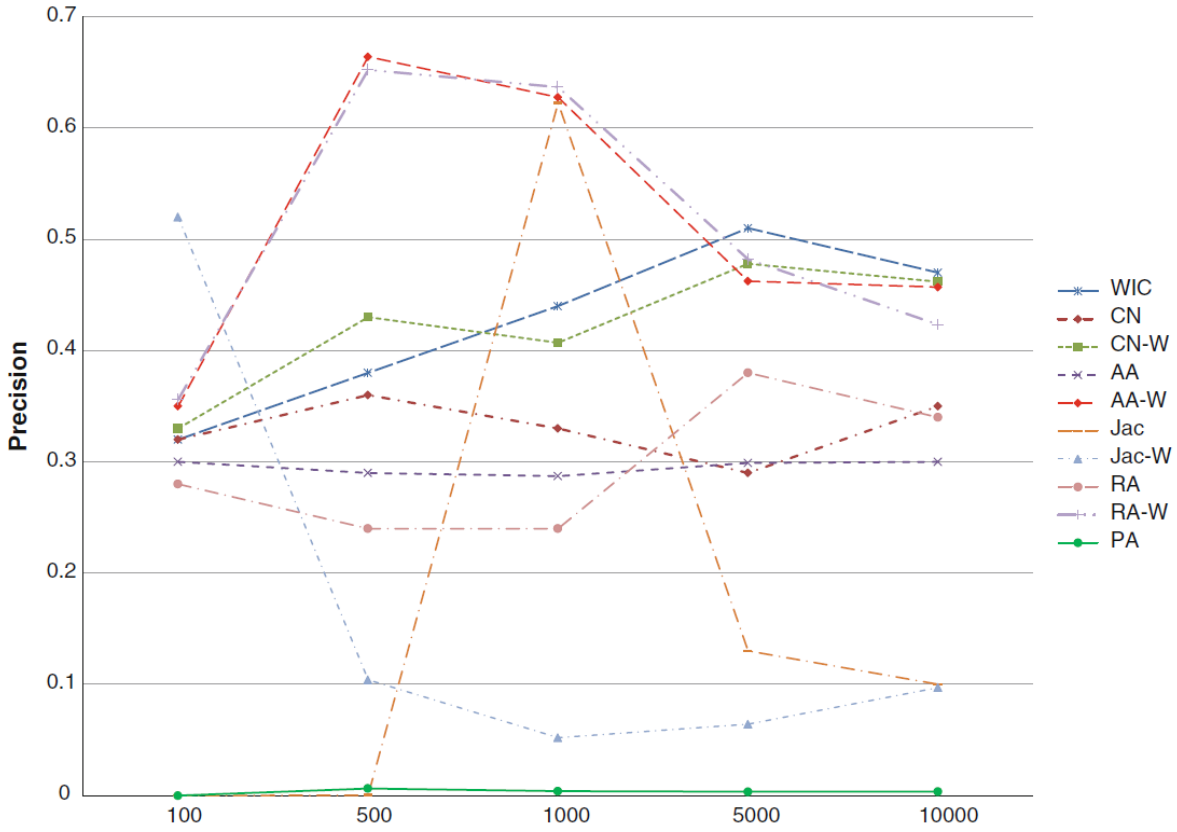


Table 2 Link prediction results measured by AUC on two subgraphs of Twitter network. The emphasized values correspond to the highest results among the evaluated measures

Graph	WIC	CN	CN-W	AA	AA-W	Jac	Jac-W	RA	RA-W	PA
Twitter 7it	0.62	0.56	0.59	0.53	0.58	0.45	0.56	0.6	0.61	0.51
Twitter 15it	0.62	0.56	0.59	0.53	0.58	0.45	0.56	0.6	0.61	0.51

Experiments and Results

Precision results on the two graphs from twitter network



Experiments and Results

Table 3 Accuracy results (in percent) on four Twitter data sets formed by feature vectors that combine different link prediction measures

Data set	J48	NB	SMO	MLP
VLocal	83.74 (0.24)	71.63 (0.28)	81.35 (0.19)	82.73 (0.25)
VGroup	82.86 (0.25)	71.70 (0.28)	80.00 (0.20)	81.88 (0.26)
VTop	83.08 (0.25)	72.12 (0.28)	80.34 (0.20)	82.01 (0.26)
VTotal	83.80 (0.24)	72.05 (0.28)	81.71 (0.18)	82.84 (0.24)

Values in parenthesis indicate the mean absolute error

Table 4 F-measure results on four Twitter data sets formed by feature vectors that combine different link prediction measures

Data set	J48	NB	SMO	MLP
VLocal	0.837	0.698	0.812	0.827
VGroup	0.829	0.699	0.798	0.819
VTop	0.831	0.703	0.801	0.820
VTotal	0.838	0.703	0.816	0.828

Overview

- Introduction
- Problem description
  - Unsupervised strategy
  - Supervised strategy
- Structural link prediction measures
- Prediction with community information
  - Label Propagation algorithm
  - WIC measure
  - W form measures
- Experiments and Results
  - Experimental setup
  - Results and analysis
- Conclusion
- Recommendation
- References

Conclusion

- The WIC measure and the W form measure for detecting the community
- For a fast community detection, LPA algorithm is used
- Considering time execution and AUC criterion, WIC performs better than other measures
- WIC and W form measures improve link prediction performance because nodes of same communities likely have similar interests or behaviours

Overview

- Introduction
- Problem description
  - Unsupervised strategy
  - Supervised strategy
- Structural link prediction measures
- Prediction with community information
  - Label Propagation algorithm
  - WIC measure
  - W form measures
- Experiments and Results
  - Experimental setup
  - Results and analysis
- Conclusion
- Recommendation
- References

Recommendation

- Link prediction for undirected networks
- Experimental analysis using supervised strategy
- What would happen with multiple links?

Future Work

- To implement an algorithm different from LPA that would reduce the execution time and make WIC and W form measures perform even better

Overview

- Introduction
- Problem description
  - Unsupervised strategy
  - Supervised strategy
- Structural link prediction measures
- Prediction with community information
  - Label Propagation algorithm
  - WIC measure
  - W form measures
- Experiments and Results
  - Experimental setup
  - Results and analysis
- Conclusion
- Recommendation
- References

References

1. Valverde-Rebaza J, de Andrade Lopes A (2012) Structural link prediction using community information on twitter. In: Computational aspects of social networks (CASoN), 2012 Fourth International Conference on, Nov 2012, pp 132–13
2. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76: 036106
3. Zhang Q-M, Lu¨ L, Wang W-Q, Zhu Y-X, Zhou T (2012) Potential theory for directed networks. CoRR abs/1202.2709
4. Valverde-Rebaza J, de Andrade Lopes A (2012) Link prediction in complex networks based on cluster information. In: Advances in artificial intelligence, SBIA 2012, 21th Brazilian symposium on artificial intelligence, ser, Vol 7589. Lecture Notes in Computer Science, Springer 92–101
5. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Dankeschön