UTS UNIVERSITY OF TECHNOLOGY SYDNEY

COMPLEX NETWORKS

Closure Coefficient in Complex Directed Networks



Mingshan Jia, UTS
mingshan.jia@student.uts.edu.au

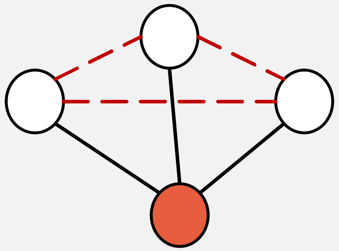
Bogdan Gabrys
UTS

Katarzyna Musial
UTS

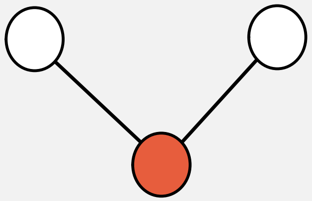
Clustering in Networks: The Local Clustering Coefficient

- A measure of cliquishness of a neighbourhood.

- The **local clustering coefficient** captures the degree to which the neighbours of a focal node connect to each other.



- The focal node serves as the **centre-node** in an open triad.



[D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks”, 1998]

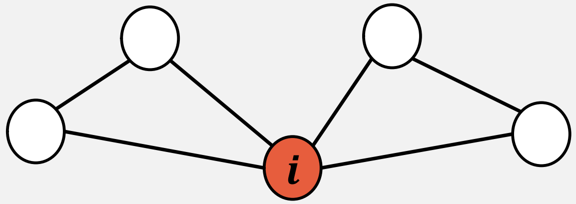
Clustering in Networks: The Local Clustering Coefficient

- Notation:
  Let $G = (V, E)$ be an undirected graph on a node set V and an edge set E
  The adjacency matrix G is denoted as $\boldsymbol{A} = \{a_{ij}\}$.
  $a_{ij} = 1$ if there is an edge between node i and node j, otherwise $a_{ij} = 0$.
  The degree of node i is denoted d_i.

- For any node $i \in V$, the **local clustering coefficient** is defined as:

$$C(i) = \frac{T(i)}{OTC(i)} = \frac{\frac{1}{2}\sum_j \sum_k a_{ij}a_{ik}a_{jk}}{\frac{1}{2}d_i(d_i - 1)}, \qquad (1)$$

where $OTC(i)$ is the number of open triads with i as the centre-node.



$$T(i) = 2$$

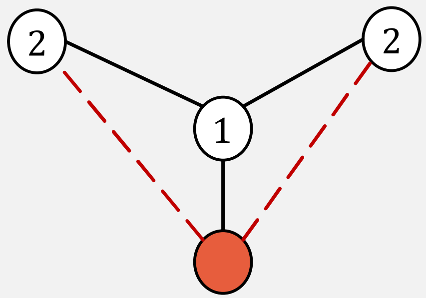
$$OTC(i) = \frac{4 * 3}{2} = 6$$

$$C(i) = 0.33$$

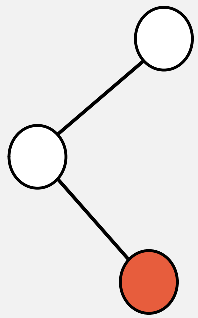
Closure Coefficient

- Another measure of clustering for undirected networks.

- The **local closure coefficient** captures the degree to which the 2-hop neighbours of a focal node connect to the focal node itself.



- The focal node serves as the **end-node** in an open triad.

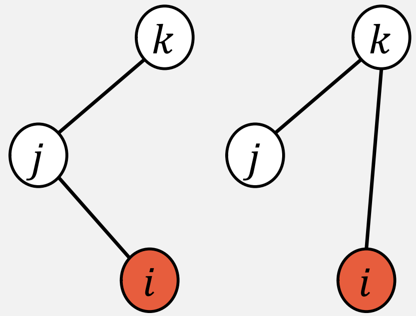


[H. Yin, A. R. Benson, and J. Leskovec, “The local closure coefficient: A new perspective on network clustering”, 2019]

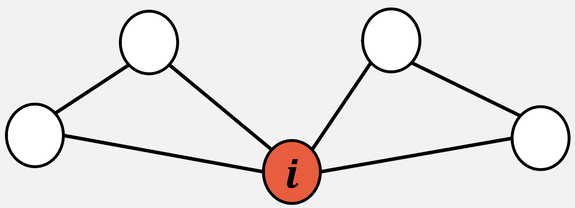
Closure Coefficient

- For any node $i \in V$, the **local closure coefficient** is defined as:

$$E(i) = \frac{2 * T(i)}{OTE(i)} = \frac{\sum_j \sum_k a_{ij} a_{ik} a_{jk}}{\sum_{j \in N(i)} (d_j - 1)} \, , \qquad (2)$$



where $OTE(i)$ is the number of open triads with i as the end-node.
$N(i)$ denotes the set of neighbours of node i.



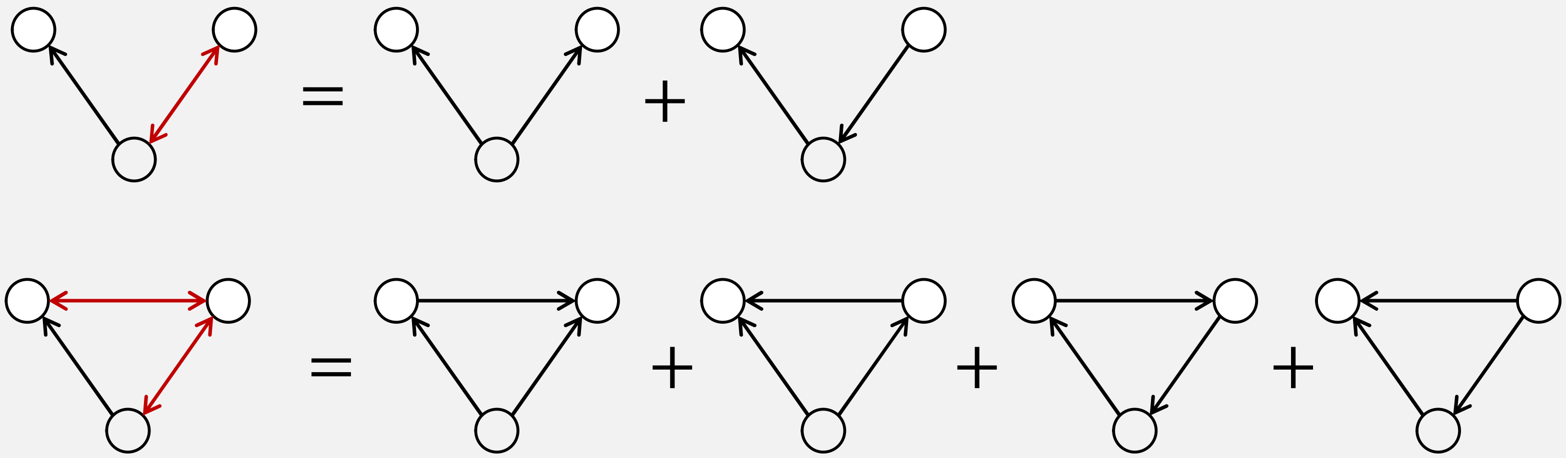
$$T(i) = 2$$

$$OTE(i) = 4$$

$$E(i) = \frac{2 * 2}{4} = 1$$

Closure Coefficient in Directed Networks

One bidirectional edge is counted as two unidirectional edges.

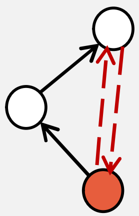


Closure Coefficient in Directed Networks

- Notation:
  Let $\boldsymbol{A} = \{a_{ij}\}$ denote the adjacency matrix of a directed graph $G^D = (V, E)$.
  $a_{ij} = 1$ if there is an edge from node i to node j, otherwise $a_{ij} = 0$.
  $N(i)$ denote the set of neighbours of node i, including both out-neighbours and in-neighbours.
  The degree of node i is denoted d_i.

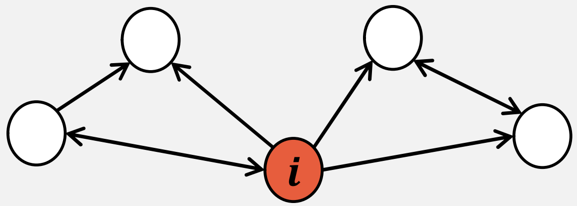
$$d_i = d_i^{out} + d_i^{in} = \sum_j a_{ij} + \sum_i a_{ji}$$

- For any node $i \in V$, the **local directed closure coefficient** is defined as:



$$\boxed{E^D(i) = \frac{2 * T^D(i)}{2 * OTE^D(i)}} = \frac{\sum_j \sum_k (a_{ij}+a_{ji})(a_{ik} + a_{ki})(a_{jk} + a_{kj})}{2 * \sum_{j \in N(i)} (a_{ij} + a_{ji})(d_i - (a_{ij} + a_{ji}))} \quad (3)$$

$$\boxed{E(i) = \frac{2 * T(i)}{OTE(i)}}$$



$$T^D(i) = 4$$

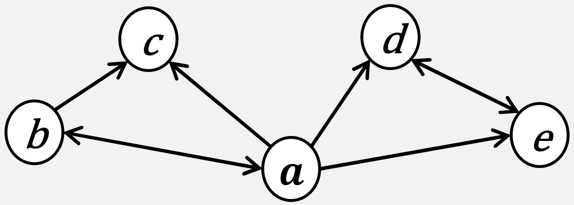
$$OTE^D(i) = 7$$

$$E^D(i) = \frac{2 * 4}{2 * 7} = 0.57$$

Average Closure Coefficient in Directed Networks

- In order to measure at the network-level, we propose the **average directed closure coefficient**. It is defined as the average of the local directed closure coefficient over all nodes:

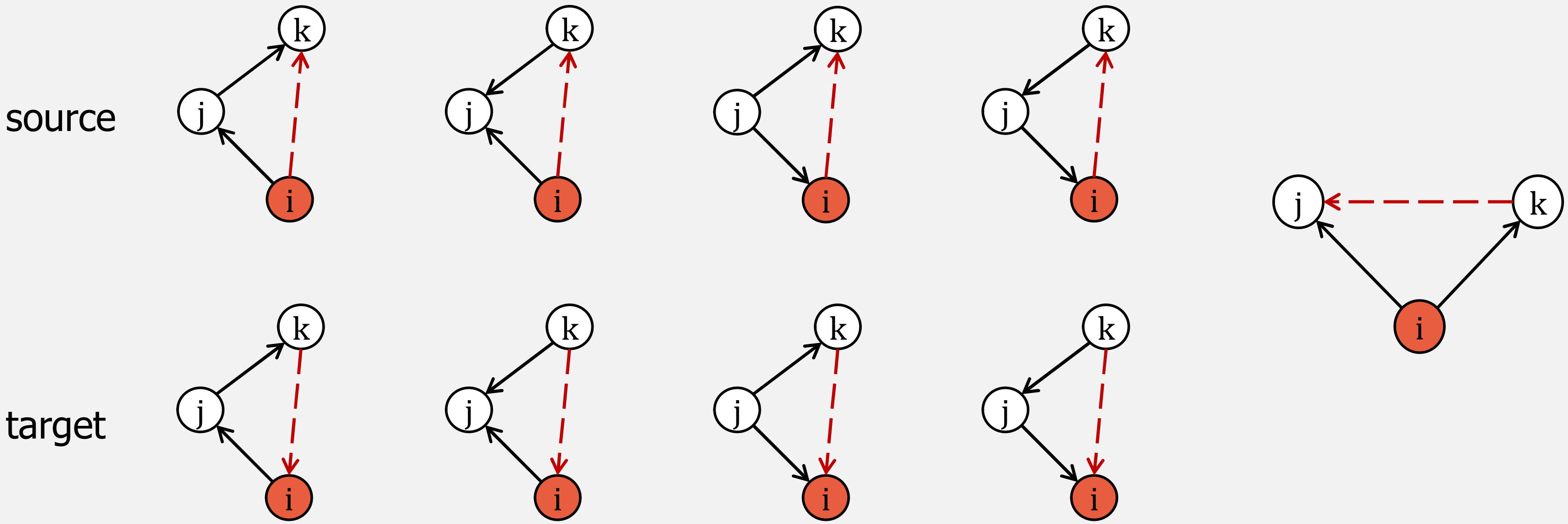
$$\overline{E^D} = \frac{1}{|V|} \sum_{i \in V} \mathrm{E}^D(i) \qquad (4)$$



$$E^D(a) = 0.57$$
$$E^D(b) = 0.29$$
$$E^D(c) = 0.33$$
$$E^D(d) = 0.33$$
$$E^D(e) = 0.33$$

$$\overline{E^D} = \frac{E^D(a) + E^D(b) + E^D(c) + E^D(d) + E^D(e)}{5} = 0.37$$

Two Types of Directed Closure Coefficient

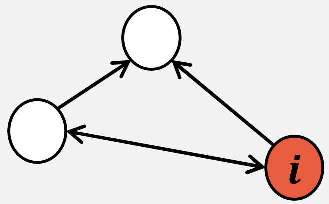


Source Closure Coefficient & Target Closure Coefficient

For a given node i in a directed network, the **source closure coefficient**, denoted $E^{src}(i)$, and the **target closure coefficient**, denoted $E^{tgt}(i)$, are defined as:

$$E^{src}(i) = \frac{T^{src}(i)}{2 * OTE^{D}(i)} = \frac{\sum_j \sum_k (a_{ij}+a_{ji})(a_{jk} + a_{kj})\, a_{ik}}{2 * \sum_{j \in N(i)} (a_{ij} + a_{ji})(d_i - (a_{ij} + a_{ji}))}\ , \quad (5)$$

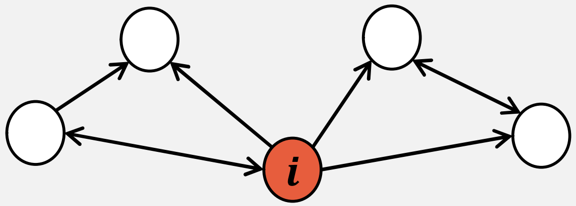
$$E^{tgt}(i) = \frac{T^{tgt}(i)}{2 * OTE^{D}(i)} = \frac{\sum_j \sum_k (a_{ij}+a_{ji})(a_{jk} + a_{kj})\, a_{ki}}{2 * \sum_{j \in N(i)} (a_{ij} + a_{ji})(d_i - (a_{ij} + a_{ji}))}\ . \quad (6)$$



$T^{src}(i) = 3$

$T^{tgt}(i) = 1$

$$E^{D}(i) = E^{src}(i) + E^{tgt}(i)$$



$T^{src}(i) = 7$

$T^{tgt}(i) = 1$

$OTE^{D}(i) = 7$

$$E^{src}(i) = \frac{7}{2 * 7} = 0.50$$

$$E^{tgt}(i) = \frac{1}{2 * 7} = 0.07$$

Closure Coefficient in Weighted Networks

Weighted undirected networks:

- Notation:
  Let $\boldsymbol{W} = \{w_{ij}\}$ denote the weight matrix of a weighted graph G^W.
  $w_{ij} \in [0, 1]$, all weights are normalized by the maximum weight.
  $N(i)$ denote the set of neighbours of node i.
  The strength of node i is denoted $s_i = \sum_j w_{ij}$.

- The **weighted closure coefficient** of node i is defined as:

$$E^W(i) = \frac{\sum_j \sum_k w_{ij} w_{ik} w_{jk}}{\sum_{j \in N(i)} w_{ij} (s_i - w_{ij})} \quad . \qquad (7)$$

- When network becomes binary, $E^W(i) = E(i)$.

Closure Coefficient in Weighted Networks

Weighted directed networks:

- Notation:
  Let $\boldsymbol{W} = \{w_{ij}\}$ denote the weight matrix of a weighted directed graph $G^{W,D}$.
  $w_{ij} \in [0, 1]$, all weights are normalized by the maximum weight.
  $N(i)$ denote the set of neighbours of node i, including both out-neighbours and in-neighbours.
  The strength of node i is denoted $s_i = \sum_j w_{ij} + \sum_j w_{ji}$.

- The **weighted directed closure coefficient** of node i is defined as:

$$E^{W,D}(i) = \frac{\sum_j \sum_k (w_{ij}+w_{ji})(w_{ik} + w_{ki})(w_{jk} + w_{kj})}{2 * \sum_{j \in N(i)} (w_{ij} + w_{ji})(s_i - (w_{ij} + w_{ji}))} \quad . \qquad (8)$$

- This definition can also be used in **weighted signed** networks (where $w_{ij} \in [-1, 1]$), only with a modified definition of $s_i = \sum_j |w_{ij}| + \sum_j |w_{ji}|$.

Computational Efficiency

- Local directed closure coefficient $E^D(i)$:

$$E^D(i) = \frac{\sum_j \sum_k (a_{ij}+a_{ji})(a_{ik}+a_{ki})(a_{jk}+a_{kj})}{2 * \sum_{j \in N(i)} (a_{ij}+a_{ji})(d_i-(a_{ij}+a_{ji}))}$$

$\boldsymbol{O(\bar{k}^2)}$

- Average directed closure coefficient $\overline{E^D}$:

$$\overline{E^D} = \frac{1}{|V|} \sum_{i \in V} \mathrm{E}^D(i)$$

$\boldsymbol{O(n \cdot \bar{k}^2)}$

Experiments: Datasets

Table 1. Statistics of datasets, showing the number of nodes ($|V|$), the number of edges ($|E|$), the average degree($\bar{k}$), the proportion of reciprocal edges (r), the average directed clustering coefficient ($\overline{C^{\mathcal{D}}}$), and the average directed closure coefficient ($\overline{E^{\mathcal{D}}}$) defined in this paper. Datasets having timestamps on edge creation are superscripted by (τ). Positively weighted networks are superscripted by ($+$), and networks having both positive and negative weights are superscripted by ($\pm$).

Network	$\lvert V\rvert$	$\lvert E\rvert$	$\bar{k}$	r	$\overline{C^{\mathcal{D}}}$	$\overline{E^{\mathcal{D}}}$
CollegeMsg$^{\tau}$	1,899	20,296	10.69	0.636	0.087	0.017
Ado-Health^{+}	2539	12,969	5.11	0.388	0.090	0.071
BTC-Alpha$^{\pm,\tau}$	3783	24,186	6.39	0.832	0.046	0.006
Wiki-Vote	7,115	104K	14.57	0.056	0.082	0.017
Epinions$^{\pm,\tau}$	132K	841K	6.38	0.308	0.085	0.010
Digg-Friends$^{\tau}$	280K	1,732K	6.19	0.212	0.075	0.008
Arxiv-HepPh	34,546	422K	12.2	0.003	0.143	0.053
US-Patent	3,775K	16,519K	4.38	0.000	0.038	0.019
AskUbuntu$^{\tau}$	79,155	199K	2.51	0.002	0.028	2e-4
StackOverflow$^{\tau}$	2,465K	16,266K	6.60	0.002	0.008	2e-4
Amazon	403K	3,387K	8.40	0.557	0.364	0.234
Google	876K	5,105K	5.83	0.307	0.370	0.097

1 communication network

2 friendship networks

3 trust networks

2 citation networks

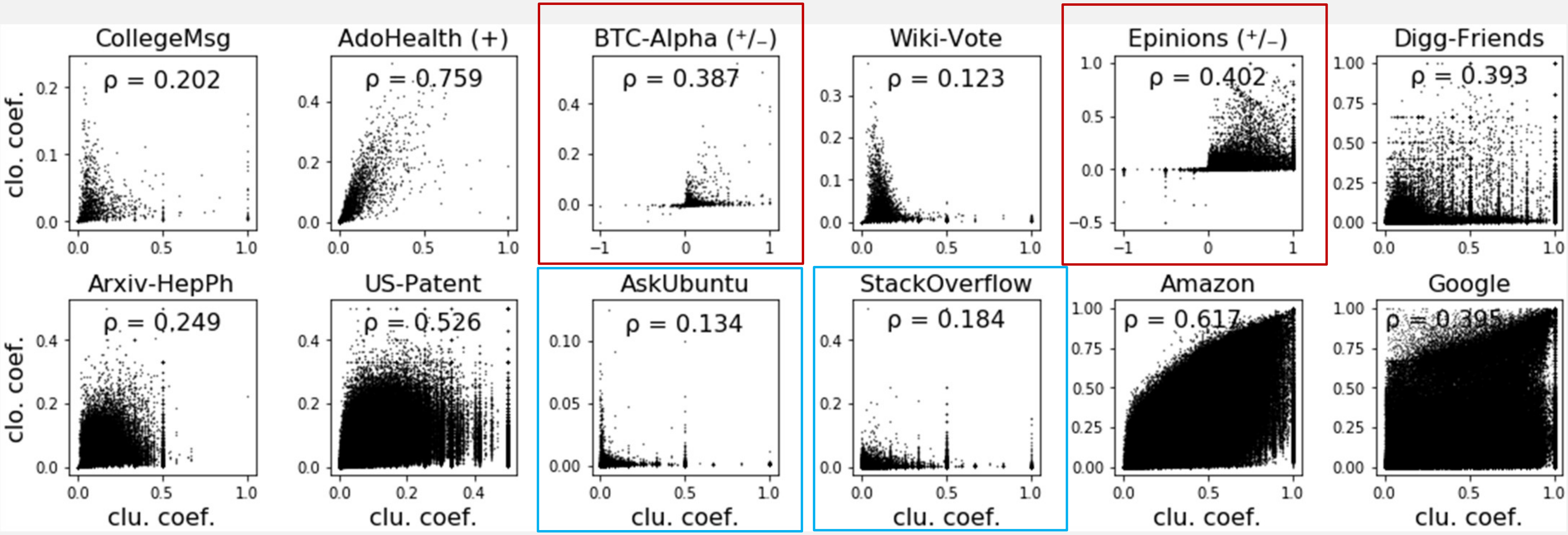
2 online Q&A networks

1 co-purchased product network

1 hyperlink network

[G. Fagiolo, "Clustering in complex directed networks", 2007]

Experiments: Correlation with Local Directed Clustering Coef.

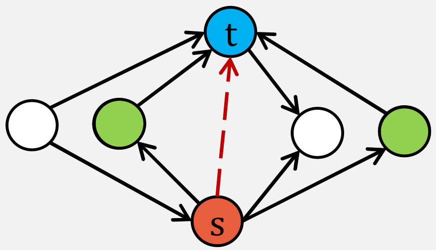


- The directed closure coefficient provides complementary information to the classic directed clustering coefficient.

Experiments: Link Prediction in Directed Networks

Classic neighbourhood based methods for undirected links:

- Common Neighbours Index (CN): $CN(x, y) = |N(x) \cap N(y)|$
- Adamic-Adar Index (AA): $AA(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$
- Resource Allocation Index (RA): $RA(s, t) = \sum_{u \in N(x) \cap N(y)} \frac{1}{|N(u)|}$



$CN(s, t) = 4$

$DiCN(s, t) = 2$

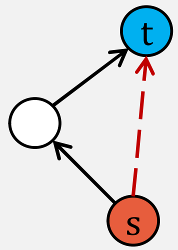
Variations of neighbourhood methods for directed links:

- Directed Common Neighbours Index (DiCN): $DiCN(s, t) = |N_{out}(s) \cap N_{in}(t)|$
- Directed Adamic-Adar Index (DiAA): $DiAA(s, t) = \sum_{u \in N_{out}(s) \cap N_{in}(t)} \frac{1}{\log |N(u)|}$
- Directed Resource Allocation Index (DiRA): $DiRA(s, t) = \sum_{u \in N_{out}(s) \cap N_{in}(t)} \frac{1}{|N(u)|}$

Experiments: Directed Closure Coefficient in Link Prediction

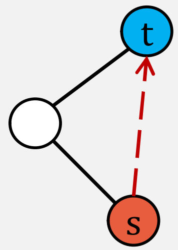
Proposed indices for directed links:

- Closure Closeness Index (CCI):

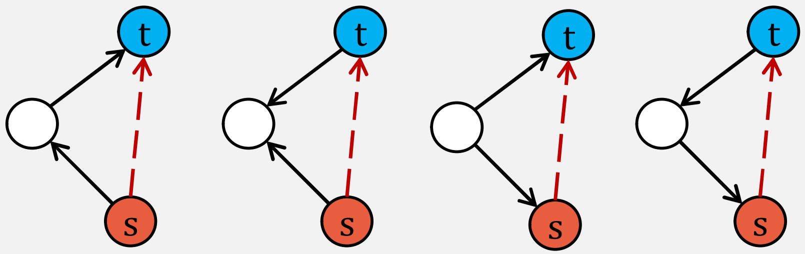


$$CCI(s,t) = |N_{out}(s) \cap N_{in}(t)| \cdot (E^{src}(s) + E^{tgt}(t))$$

- Extra Closure Closeness Index (ECCI):

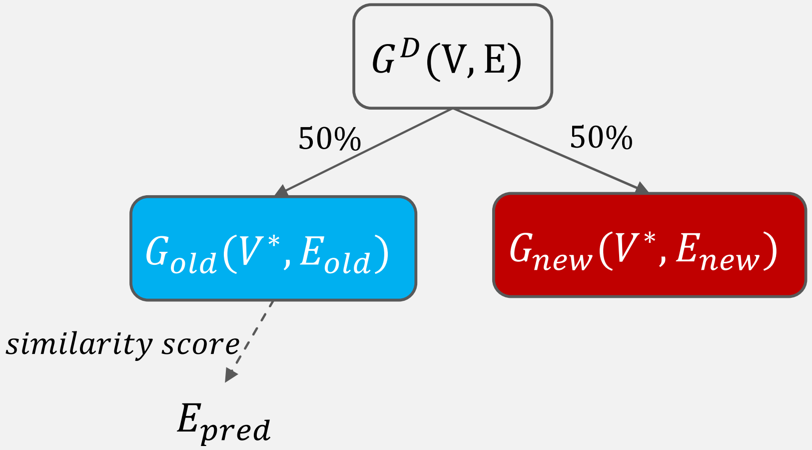


$$ECCI(s,t) = N\ (s) \cap N(t)| \cdot (E^{src}(s) + E^{tgt}(t))$$



Experiments: Directed Closure Coefficient in Link Prediction

Setup:



$$precision = \frac{|E_{pred} \cap E_{new}|}{|E_{new}|}$$

- Sampling:

In very large networks ($n > 10K$): we perform a randomised breadth first search sampling of 5K nodes on G^D, and repeat it many times according to the size of the dataset.

Experiments: Directed Closure Coefficient in Link Prediction

Table 2. Performance comparison of six methods on link prediction in directed networks (Precision %). RP (second column) gives the probability that a random prediction is correct. The best performance in each network is in bold type.

Network	RP	DiCN	DiAA	DiRA	CCI	ECCI
CollegeMsg$^{\tau}$	0.30	2.546	2.763	3.533	3.395	**3.730**
Ado-Health	0.10	8.404	8.406	8.304	10.23	**11.07**
BTC-Alpha$^{\tau}$	0.05	8.588	**9.269**	7.313	8.418	9.226
Wiki-Vote	0.15	21.96	22.51	20.32	**22.55**	19.08
Epinions$^{\tau}$	0.37	3.613	3.662	3.531	3.490	**5.106**
Digg-Friends$^{\tau}$	0.33	6.649	6.709	6.685	**7.135**	5.569
Arxiv-HepPh	0.16	20.35	**21.51**	20.72	20.07	21.49
US-Patent	0.04	9.787	10.14	9.987	**11.67**	11.31
AskUbuntu$^{\tau}$	0.03	4.100	4.912	4.163	**5.412**	4.697
StackOverflow$^{\tau}$	0.16	7.433	8.129	7.472	**8.792**	6.388
Amazon	0.06	23.71	27.94	27.43	26.76	**29.46**
Google	1.19	44.48	**52.32**	50.29	49.39	46.24

Conclusion

- The directed closure coefficient
- The closure coefficient in weighted networks
- The source closure coefficient & the target closure coefficient
- Two neighbourhood based indices for directed link prediction

- The directed closure coefficient provides complementary information to the classic directed clustering coefficient.
- Including closure coefficient leads to significant improvement in directed link prediction.

$$E_i = \frac{2L_i}{\sum_{j\in N(i)}(d_j - 1)}$$