Corrected overlap weight and clustering coefficient

Vladimir Batagelj

Institute of Mathematics, Physics and Mechanics,
Department of Theoretical Computer Science,
Jadranska 19, 1 000 Ljubljana, Slovenia
and
University of Primorska, Andrej Marušič Institute,
Muzejski trg 2, Koper, Slovenia
e-mail: `vladimir.batagelj@uni-lj.si`

August 20, 2016 / 04 : 16

Abstract

We discuss two well known network measures: the overlap weight of an edge and the clustering coefficient of a node. For both of them it turns out that they are not very useful for data analytic task to identify important elements (nodes or links) of a given network. The reason for this is that they attain the largest values on complete subgraphs of relatively small size that are more probable to appear in a network than that of larger size. We show how the definitions of these measures can be corrected in such a way that they give the expected results. We illustrate the proposed corrected measures by applying them on the US Airports network using the program Pajek.

Keywords: social network analysis, importance measure, triangular weight, overlap weight, clustering coefficient

1   Introduction

1.1   Network element importance measures

To identify important / interesting elements (nodes, links) in a network we often try to express our intuition about their importantance using an appropriate measure (index, weight) following the scheme

larger is the measure value of an element, more important / interesting is this element.

Too often, in analysis of networks, researchers uncritically pick some measure from the literature (degrees, closeness, betweenness, hubs and authorities, clustering coefficient, etc. (Wasserman and Faust, 1995; Todeschini and Consonni, 2009)) and apply it to their network.

In this paper we discuss two well known network measures: the overlap weight of an edge (Onnela et al., 2007) and the clustering coefficient of a node (Holland and Leinhardt, 1971; Watts and Strogatz, 1998).

For both of them it turns out that they are not very useful for data analytic task to identify important elements of a given network. The reason for this is that they attain their largest values on complete subgraphs of relatively small size – they are more probable to appear in a network than that of larger size. We show how their definitions can be corrected in such a way that they give the expected results. We illustrate the proposed corrected measures by applying them on the US Airports network using the program Pajek.

2 Overlap weight

2.1 Overlap weight

A direct measure of the overlap of an edge $e = (u : v) \in \mathcal{E}$ in an undirected simple graph $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ is the number of common neighbors of its end-nodes u and v (see Figure 1). It is equal to $t(e)$ – the *number of triangles* (cycles of length 3) to which the edge e belongs. There are two problems with this measure:

- it is not normalized (bounded to $[0, 1]$);

- it does not consider the 'potentiality' of nodes u and v – there are

$$\min(\deg(u), \deg(v)) - 1 - t(e)$$

nodes in the smaller set of neighbors that are not in the other set of neighbors.
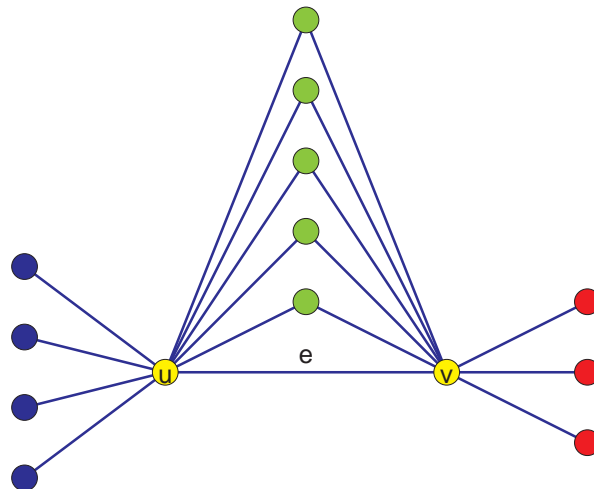

Figure 1: Neighbors of $e(u : v)$

Two simple normalizations are:

$$\frac{t(e)}{n - 2} \quad \text{or} \quad \frac{t(e)}{\mu}$$

where $n = |\mathcal{V}|$ is the number of nodes, and $\mu = \max_{e \in \mathcal{E}} t(e)$ is the maximum number of triangles on an edge in the graph $\mathbf{G}$.

The (topological) *overlap weight* of an edge $e = (u : v) \in \mathcal{E}$ considers also the degrees of edge's end-nodes and is defined as

$$o(e) = \frac{t(e)}{(\deg(u) - 1) + (\deg(v) - 1) - t(e)}$$

In the case $\deg(u) = \deg(v) = 1$ we set $o(e) = 0$. It somehow resolves both problems.

The overlap weight is essentially a Jaccard similarity index (Wikipedia, 2016)

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

for $X = N(u) \setminus \{v\}$ and $Y = N(v) \setminus \{u\}$ where $N(z)$ is the set of neighbors of a node z. In this case we have $|X \cap Y| = t(e)$ and

$$|X \cup Y| = |X| + |Y| - |X \cap Y| = (\deg(u) - 1) + (\deg(v) - 1) - t(e).$$

Note also that $h(X, Y) = 1 - J(X, Y) = \frac{|X \oplus Y|}{|X \cup Y|}$ is the normalized Hamming distance (Wikipedia, 2016). The operation $\oplus$ denotes the symmetric difference $X \oplus Y = (X \cup Y) \setminus (X \cap Y)$.

Another normalized overlap measure is the *overlap index* (Wikipedia, 2016)

$$O(X, Y) = \frac{|X \cap Y|}{\max(|X|, |Y|)}.$$

Both measures J and O, applied to networks, have some nice properties. For example: a pair of nodes u and v are structurally equivalent iff $J(X, Y) = O(X, Y) = 1$. Therefore the overlap weight measures a substitutiability of one edge's end-node by the other.

Introducing two auxiliary quantities

$$m(e) = \min(\deg(u), \deg(v)) - 1 \quad \text{and} \quad M(e) = \max(\deg(u), \deg(v)) - 1$$

we can rewrite the definiton of the overlap weight

$$o(e) = \frac{t(e)}{m(e) + M(e) - t(e)}, \quad M(e) > 0$$

and if $M(e) = 0$ then $o(e) = 0$.

For every edge $e \in \mathcal{E}$ it holds $0 \le t(e) \le m(e) \le M(e)$. Therefore

$$m(e) + M(e) - t(e) \ge t(e) + t(e) - t(e) = t(e)$$

showing that $0 \le o(e) \le 1$.

The value $o(e) = 1$ is attained exactly in the case when $M(e) = t(e)$; and the value $o(e) = 0$ exactly when $t(e) = 0$.

In simple directed graphs without loops different types of triangles exist over an arc $a(u, v)$. We can define overlap weights for each type. For example: the *transitive overlap weight*

$$o_t(a) = \frac{t_t(a)}{(\text{outdeg}(u) - 1) + (\text{indeg}(v) - 1) - t_t(a)}$$
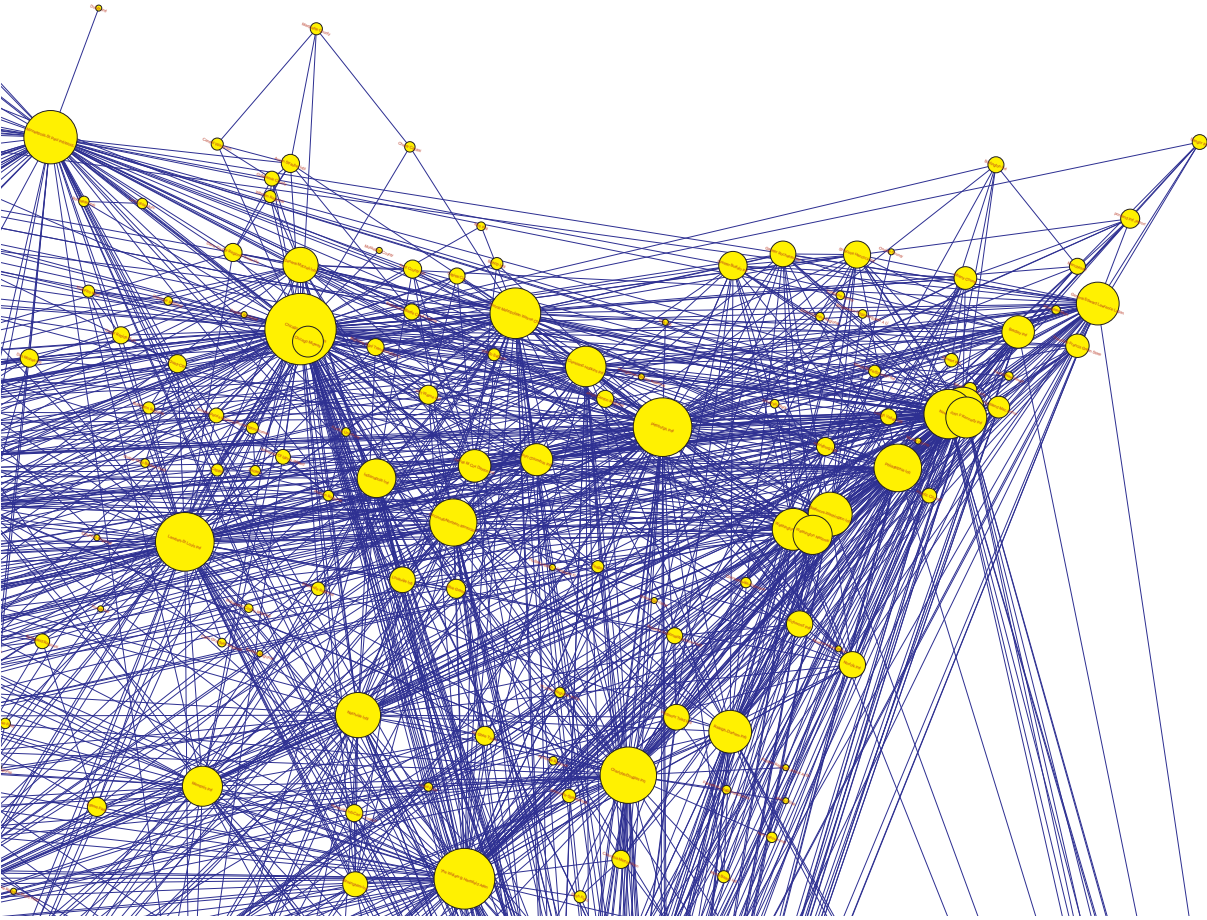
3

Figure 2: US Airports 1997 network, a North-East cut-out

and the *cyclic overlap weight*

$$o_c(a) = \frac{t_c(a)}{\mathrm{indeg}(u) + \mathrm{outdeg}(v) - t_c(a)}$$

where $t_t(a)$ and $t_c(a)$ are number of transitive / cyclic triangles containing the arc a. In this paper we will limit our discussion to overlap weights in undirected graphs.

2.2 US Airports links with the largest overlap weight

Let us apply the overlap weight to the network of US Airports 1997 (see Figure 2). It consists of 332 airports and 2126 edges among them. There is an edge linking a pair of airports iff in the year 1997 there was a flight company providing flights between those two airports.

The size of a circle representing an airport in Figure 2 is proportional to its degree – the number of airports linked to it. The airports with the largest degree are:
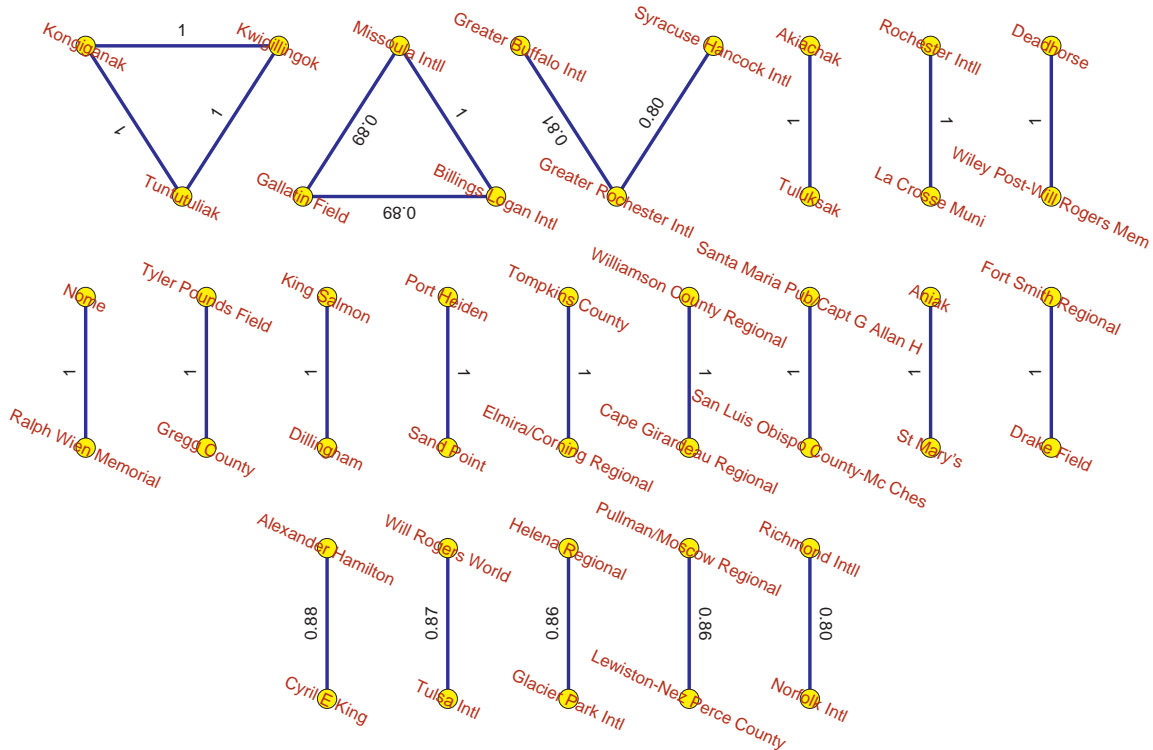
Figure 3: Edges with the largest overlap – cut at 0.8

airport	deg
Chicago O'hare Intl	139
Dallas/Fort Worth Intl	118
The William B Hartsfield Atlanta	101
Lambert-St Louis Intl	94
Pittsburgh Intl	94

For the overlap weight the edge cut at level 0.8 (a subnetwork of all edges with overlap weight at least 0.8) is presented in Figure 3. It consists of two triangles, a path of length 2, and 17 separate edges.

A tetrahedron (Kwigillingok, Kongiganak,Tuntutuliak, Bethel), see Figure 4, gives the first triangle in Figure 3 – attached with the node Bethel to the rest of network.

From this example we see that in real-life networks edges with the largest overlap weight tend to be edges with relatively small degrees in their end-nodes ($o(e) = 1$ implies $\deg(u) = \deg(v) = t(e) + 1$). Because of this the overlap weight is not very useful for data analytic tasks in searching for important elements of a given network. We would like to emphasize here that there are many applications in which overlap weight proves to be useful and appropriate; we question only its appropriateness for determining the most overlaped edges. We will try to improve the overlap weight definition to better suit the data analytic goals.
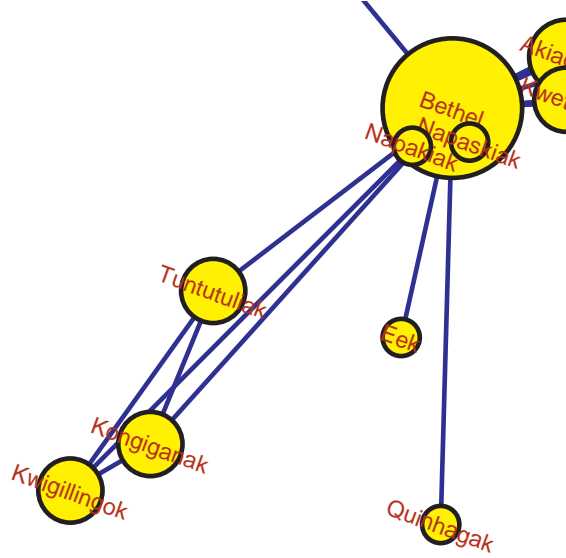
Figure 4: Zoom in

2.3 Corrected overlap weight

We define a *corrected overlap weight* as

$$o'(e) = \frac{t(e)}{\mu + M(e) - t(e)}$$

By the definiton of μ for every $e \in \mathcal{E}$ it holds $t(e) \le \mu$. Since $M(e) - t(e) \ge 0$ also $\mu + M(e) - t(e) \ge \mu$ and therefore $0 \le o'(e) \le 1$. $o'(e) = 0$ exactly when $t(e) = 0$, and $o'(e) = 1$ exactly when $\mu = M(e) = t(e)$.

2.4 US Airports 1997 links with the largest corrected overlap weight

For the US Airports 1997 network we get $\mu = 80$. For the corrected overlap weight the edge cut at level 0.5 is presented in Figure 5. Six links with the largest triangular weights are given in Table 1.

Table 1: Largest triangular weights in US Airports 1997 network

u	v	$t(e)$	$d(u)$	$d(v)$	$o'(e)$
The W B Hartsfield Atlanta	Charlotte/Douglas Intl	76	101	87	0.73077
The W B Hartsfield Atlanta	Dallas/Fort Worth Intl	73	101	118	0.58871
Chicago O'hare Intl	Pittsburgh Intll	80	139	94	0.57971
Chicago O'hare Intl	Lambert-St Louis Intl	80	139	94	0.57971
Chicago O'hare Intl	Dallas/Fort Worth Intl	78	118	139	0.55714
Chicago O'hare Intl	The W B Hartsfield Atlanta	77	101	139	0.54610

In Figure 6 all the neighbors of end-nodes WB Hartsfield Atlanta and Charlotte/Douglas Intl of the link with the largest corrected overlap weight value are presented. They have 76 common

Figure 5: US Airports 1997 links with the largest corrected overlap weight, cut at 0.5

Figure 6: US Airports links o'(WB Hartsfield Atlanta, Charlotte/Douglas Intl) = 0.7308
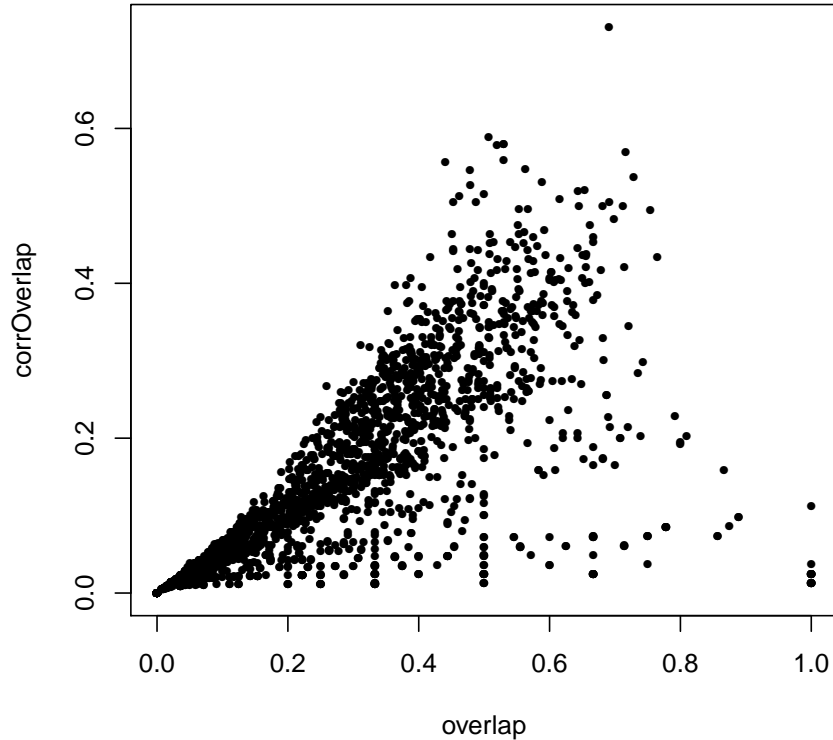
Figure 7: Comparison (overlap, corrected overlap)

(triangular) neighbors. The node WB Hartsfield Atlanta has 11 and the node Charlotte/Douglas Intl has 25 additional neighbors. Note (see Table 1) that there are some links with higher triangular weight, but also with much higher number of additional neighbors – therefore with smaller corrected overlap weights.

2.5 Comparisons

In Figure 7 the set $\{(o(e), o'(e)) : e \in \mathcal{E}\}$ is displayed for the US Airports 1997 network. For most edges it holds $o'(e) \leq o(e)$. It is easy to see that $o(e) < o'(e) \Leftrightarrow \mu < m(e)$. Edges with the overlap value $o(e) > 0.8$ have the corrected overlap weight $o'(e) < 0.2$.

In Figure 8 the sets $\{(m(e), o(e)) : e \in \mathcal{E}\}$ and $\{(m(e), o'(e)) : e \in \mathcal{E}\}$ are displayed for the US Airports 1997 network. With increasing the $m(e)$ the corresponding overlap weight $o(e)$ is decreasing; and the corresponding corrected overlap weight $o'(e)$ is also increasing.

We can observe similar tendencies if we compare both weights with respect to the number of triangles $t(e)$ (see Figure 9).
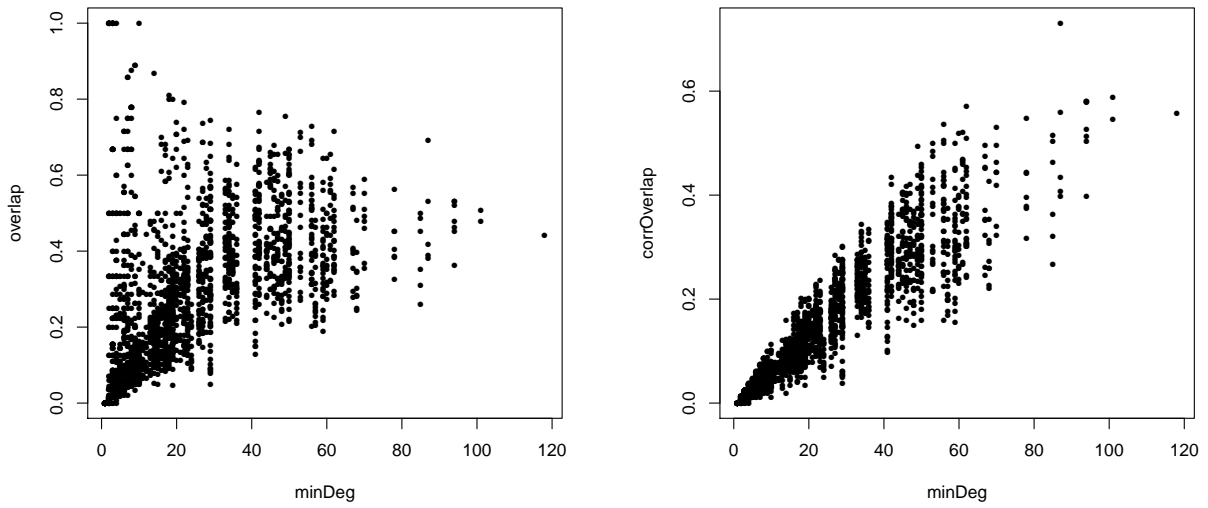
8

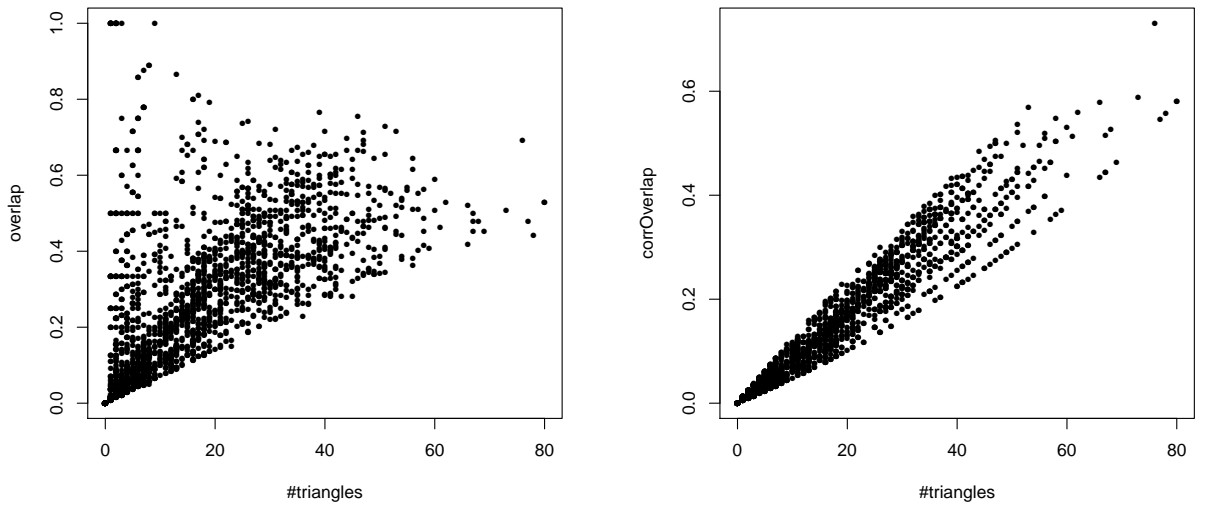Figure 8: Comparison – minDeg(e)

Figure 9: Comparison – # of triangles

3 Clustering coefficient

3.1 Clustering coefficient

For a node $u \in \mathcal{V}$ in an undirected simple graph $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ its (local) clustering coefficient (Wikipedia, 2016) is measuring a local density in the node u and is defined as a proportion of the number of existing edges between u's neighbors to the number of all possible edges between u's neighbors

$$cc(u) = \frac{|\mathcal{E}(N(u))|}{|\mathcal{E}(K_{\deg(u)})|} = \frac{2 \cdot E(u)}{\deg(u) \cdot (\deg(u) - 1)}, \quad \deg(u) > 1$$

where $E(u) = |\mathcal{E}(N(u))|$. If $\deg(u) \leq 1$ then $cc(u) = 0$.

It is easy to see that

$$E(u) = \frac{1}{2} \sum_{e \in S(u)} t(e)$$

where $S(u) = \{e(u : v) : e \in \mathcal{E}\}$ is the star in node u.

It holds $0 \leq cc(u) \leq 1$. $cc(u) = 1$ exactly when $\mathcal{E}(N(u))$ is isomorphic to $K_{\deg(u)}$ – a complete graph on $\deg(u)$ nodes. Therefore it seems that the clustering coefficient could be used to identify nodes with the densest neighborhoods.

The notion of clustering coefficient can be extended also to simple directed graphs (with loops).

3.2 US Airports links with the largest clustering coefficient

Let us apply also the clustering coefficient to the US Airports 1997 network.

In Table 2 airports with the clustering coefficient equal to 1 and the degree at least 4 are listed. There are 28 additional such airports with a degree 3, and 38 with a degree 2.

Again we see that the clustering coefficient attains its largest value in nodes with relatively small degree. The probability that we get a complete subgraph on $N(u)$ is decreasing very fast with increasing of $\deg(u)$.

Table 2: US Airports 1997 links with clustering coefficient = 1

n	deg	airport	n	deg	airport
1	7	Lehigh Valley Intll	8	4	Gunnison County
2	5	Evansville Regional	9	4	Aspen-Pitkin Co/Sardy Field
3	5	Stewart Int'l	10	4	Hector Intll
4	5	Rio Grande Valley Intl	11	4	Burlington Regional
5	5	Tallahassee Regional	12	4	Rafael Hernandez
6	4	Myrtle Beach Intl	13	4	Wilkes-Barre/Scranton Intl
7	4	Bishop Intll	14	4	Toledo Express

Table 3: US Airports 1997 nodes with the largest corrected clustering coefficient

Rank	Value	deg	Id
1	0.3739	45	Cleveland-Hopkins Intl
2	0.3700	50	General Edward Lawrence Logan
3	0.3688	56	Orlando Intl
4	0.3595	42	Tampa Intl
5	0.3488	61	Cincinnati/Northern Kentucky Intl
6	0.3457	70	Detroit Metropolitan Wayne County
7	0.3455	67	Newark Intl
8	0.3429	53	Baltimore-Washington Intl
9	0.3415	47	Miami Intl
10	0.3405	42	Washington National
11	0.3379	56	Nashville Intll
12	0.3359	46	John F Kennedy Intl
13	0.3347	62	Philadelphia Intl
14	0.3335	41	Indianapolis Intl
15	0.3335	50	La Guardia

3.3 Corrected clustering coefficient

To get a corrected version of the clustering coefficient we proposed in Pajek (De Nooy et al., 2011) to replace $\deg(u)$ in the denominator with $\Delta = \max_{v \in \mathcal{V}} \deg(v)$. In this paper we propose another solution – we replace $\deg(u) - 1$ with μ:

$$cc'(u) = \frac{2 \cdot E(u)}{\mu \cdot \deg(u)}, \quad \deg(u) > 0$$

Note that, if $\Delta > 0$ then $\mu < \Delta$.

To show that $0 \leq cc'(u) \leq 1$ we have to consider two cases:

a. $\deg(u) \geq \mu$: then for $v \in N(u)$ we have $\deg_{N(u)}(v) \leq \mu$ and therefore

$$2 \cdot E(u) = \sum_{v \in N(u)} \deg_{N(u)}(v) \leq \sum_{v \in N(u)} \mu = \mu \cdot \deg(u)$$

b. $\deg(u) < \mu$: then $\deg(u) - 1 \leq \mu$ and therefore

$$2 \cdot E(u) \leq \deg(u) \cdot (\deg(u) - 1) \leq \mu \cdot \deg(u)$$

The value $cc'(u) = 1$ is attained in the case a) on a μ-core, and in the case b) on $K_{\mu+1}$.

3.4 US Airports nodes with the largest corrected clustering coefficient

In Table 3 US Airports with the largest corrected clustering coefficient are listed. The largest value 0.3739 is attained for Cleveland-Hopkins Intl airport. In Figure 10 the adjacency matrix
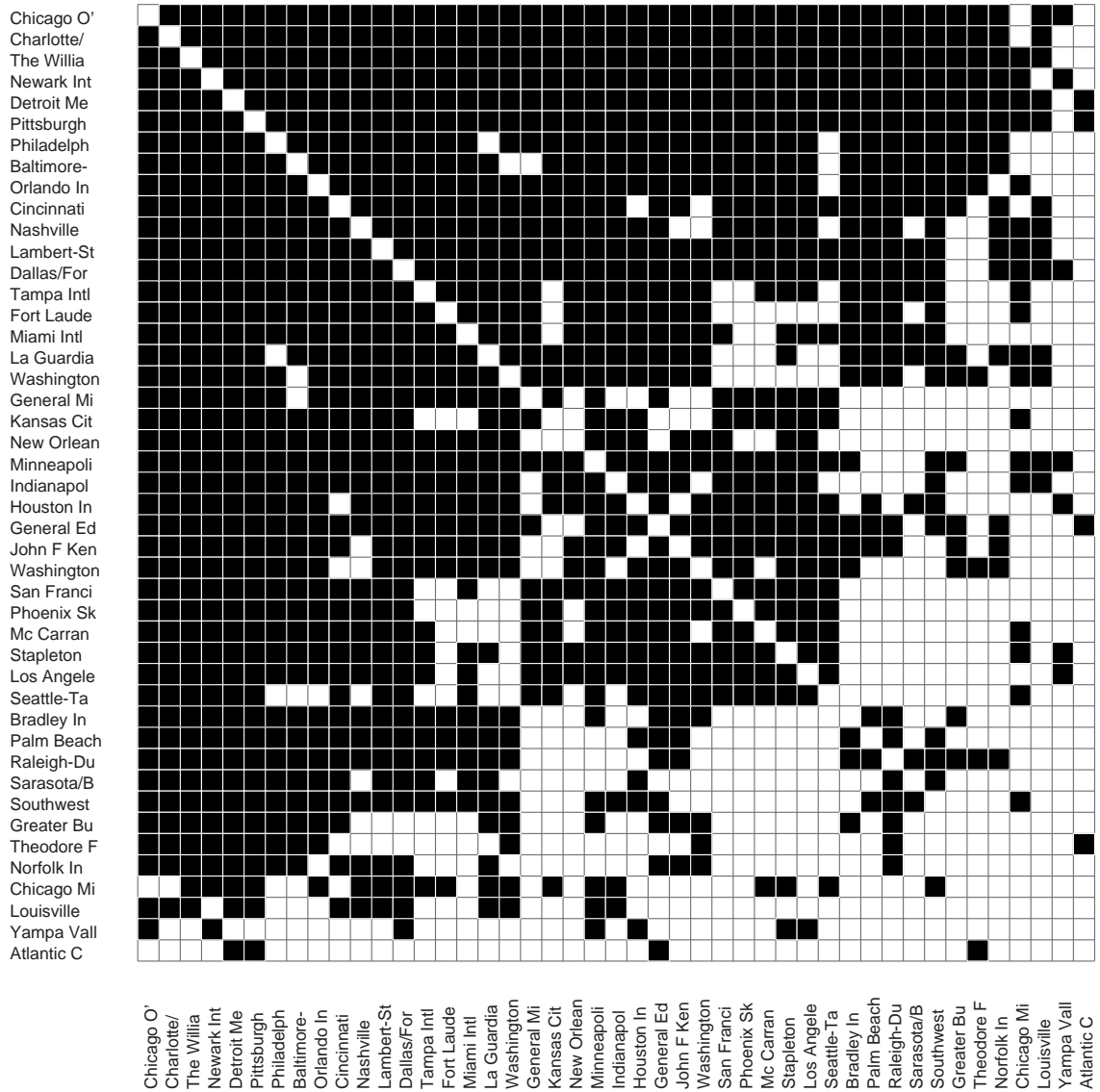
Figure 10: Links among Cleveland-Hopkins Intl neighbors

of a subnetwork on its 45 neighbors is presented. The subnetwork is relatively complete. A small value of corrected clustering coefficient is due to relatively small $\deg = 45$ with respect to $\mu = 80$.

3.5 Comparisons

In Figure 11 the set $\{(cc(e), cc'(e)) : e \in \mathcal{E}\}$ is displayed for the US Airports 1997 network. The correlation between both coefficients is very small. An important observation is that edges with the largest value of the clustering coefficient have relatively small values of the corrected clustering coefficient. We also see that the number of edges in a node's neighborhood is almost functionally dependent on its degree.

From Figure 12 we see that the clustering coefficient is decreasing with the increasing de-
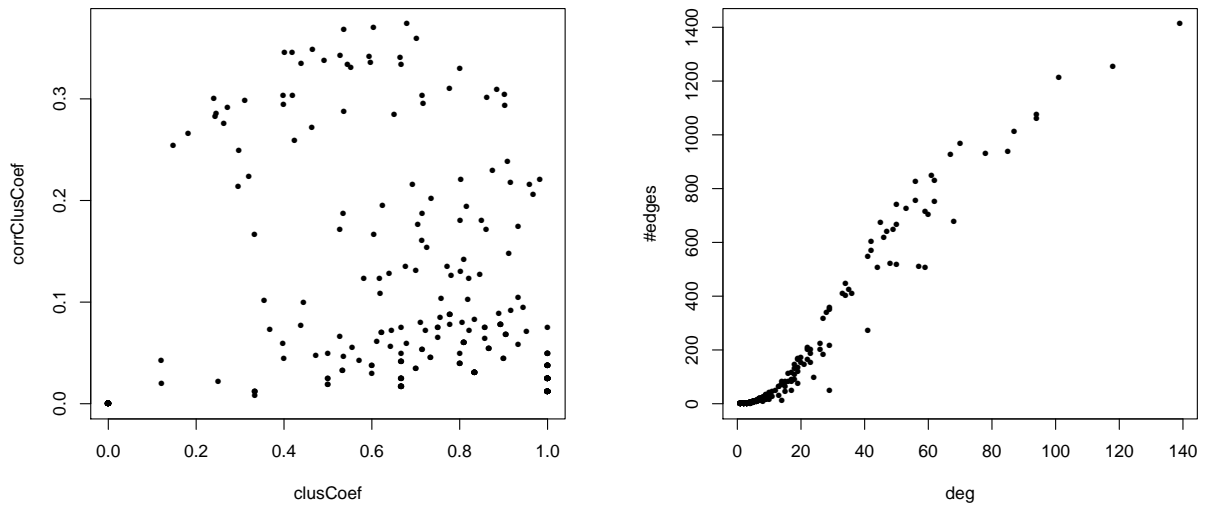
Figure 11: Comparison – ordinary and corrected clustering coefficients; degrees and number of edges
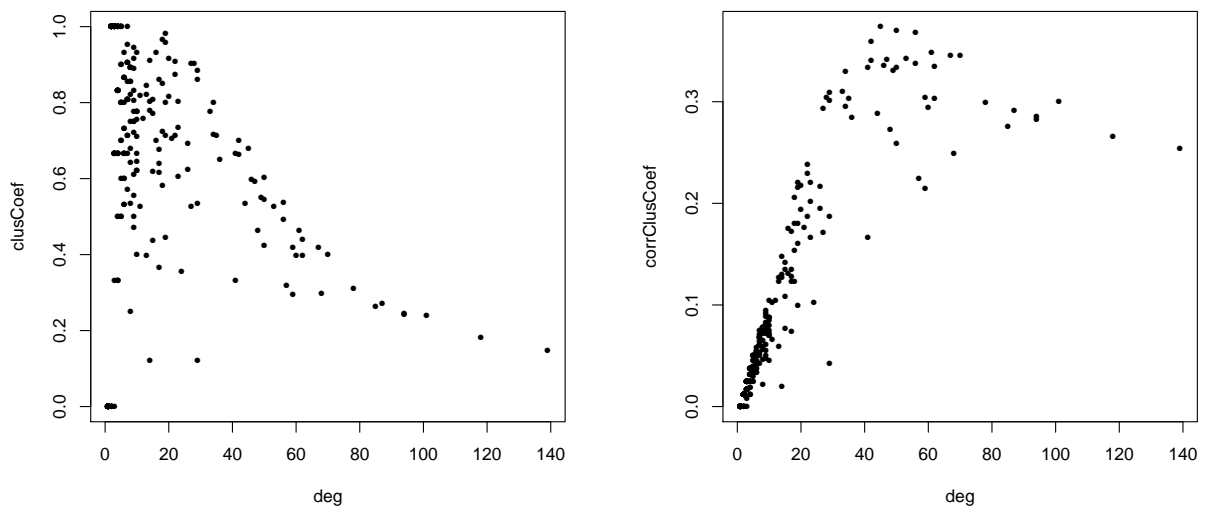

Figure 12: Comparison – degrees

gree. Nodes with large degree have small values of clustering coefficient. The values of corrected clustering coefficient are large for nodes of large degree.

4 Conclusions

In the paper we showed that two network measures, the overlap weight and clustering coefficient, are not suitable for data analytic task of determining of important elements in a given network. We proposed corrected versions of these two measures that give expected results.

Because $\mu \leq \Delta$ we can replace in the corrected measures μ with Δ. Its advantage is that it can be easier computed; but the corresponding corrected index is less 'sensitive'.

Acknowledgments

References

Batagelj, V. (2016), Corrected. https://github.com/bavla/corrected .

De Nooy, W., Mrvar, A., and Batagelj, V. (2011): Exploratory Social Network Analysis with Pajek; Revised and Expanded Second Edition. Structural Analysis in the Social Sciences, Cambridge University Press.

Holland, P.W. and Leinhardt, S. (1971). Transitivity in structural models of small groups. Comparative Group Studies 2: 107–124.

Onnela, J.P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., Kertesz, J., Barabasi, A.L. (2007). Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences 104(18), 7332.

Todeschini, R. and Consonni, V. (2009). Molecular Descriptors for Chemoinformatics, 2nd ed. Wiley-VCH, Weinheim.

Wasserman, S. and Faust, K. (1995). Social Network Analysis Methods and Applications. Structural Analysis in the Social Sciences, Cambridge University Press.

Watts, D.J. and Strogatz, S. (June 1998). Collective dynamics of 'small-world' networks. Nature 393 (6684): 440–442.

Wikipedia (2016):

Clustering coefficient: https://en.wikipedia.org/wiki/Clustering_coefficient,
Overlap coefficient: https://en.wikipedia.org/wiki/Overlap_coefficient,
Hamming distance: https://en.wikipedia.org/wiki/Hamming_distance,
Jaccard index: https://en.wikipedia.org/wiki/Jaccard_index.