Corrected network measures

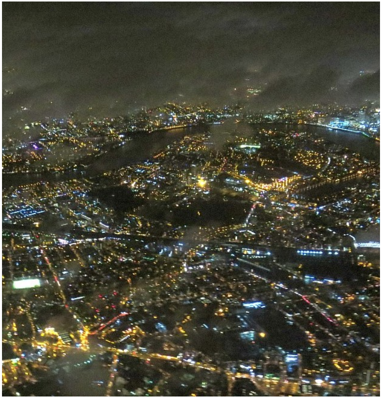
Vladimir Batagelj

IMFM Ljubljana and IAM UP Koper

CMStatistics (ERCIM) 2015

Senate House, University of London – December 12-14, 2015

Outline



Vladimir Batagelj:
vladimir.batagelj@fmf.uni-lj.si

Current version of slides (December 16, 2015, 11 : 05):
http://vlado.fmf.uni-lj.si/pub/slides/ercim15.pdf

Network element importance measures

To identify important / interesting elements (nodes, links) in a network we often try to express our intuition about important / interesting element 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.

Network element importance measures

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 the 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.

Overlap weight – definition

The (topological) *overlap weight* of an edge $e = (u : v) \in \mathcal{E}$ in an undirected simple graph $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ is defined as

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

where $t(e)$ is the number of triangles (cycles of length 3) to which the edge e belongs. In the case $\deg(u) = \deg(v) = 1$ we set $o(e) = 0$. 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

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

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

Overlap weight – properties

It holds

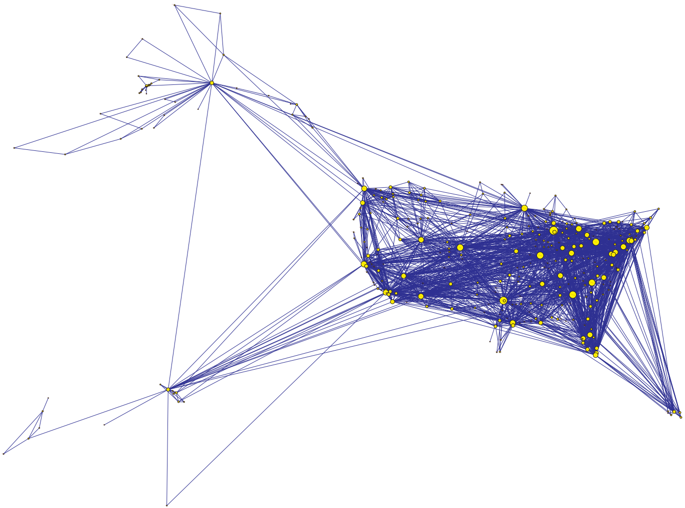
$$0 \leq t(e) \leq m(e) \leq M(e).$$

Therefore

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

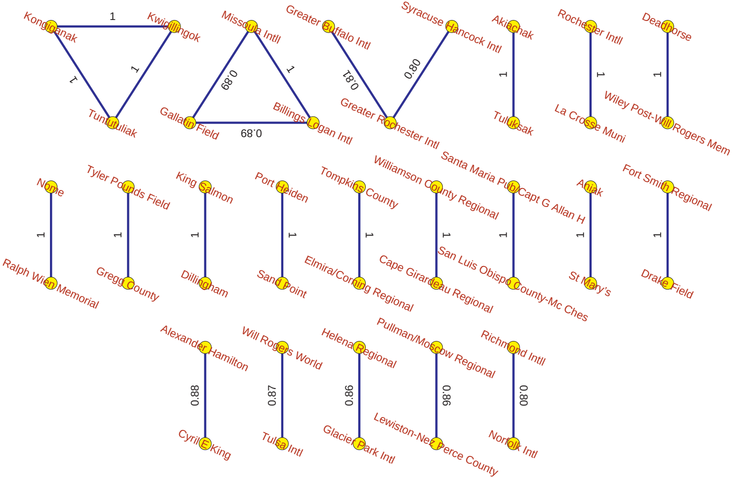
showing that $0 \leq o(e) \leq 1$.
The value $o(e) = 1$ is attained exactly in the case when $m(e) = M(e) = t(e)$; and the value $o(e) = 0$ exactly when $t(e) = 0$.

US Airports links 1997

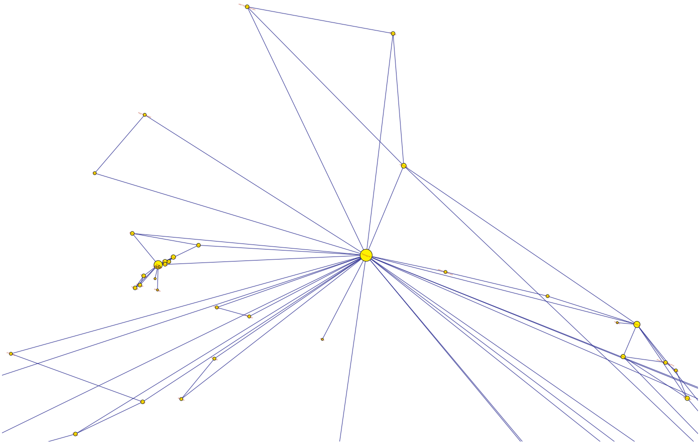


Edges with the largest overlap

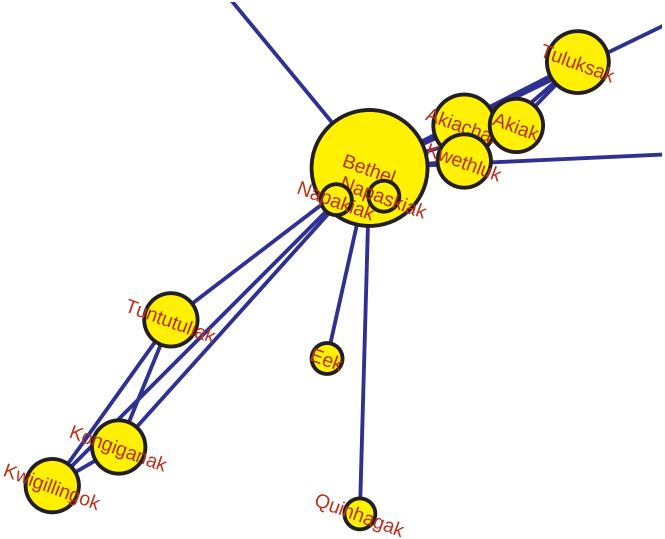
cut at 0.8



Zoom in



Zoom in


Tuluksak
Akiachak
Akiak
Kwethluk
Bethel
Napaskiak
Napakiak
Tuntutuliak
Eek
Kongiganak
Kwigillingok
Quinhagak

Observation

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. Because of this the overlap weight is not very useful for data analytic tasks in searching for important elements of a given network. We can try to improve the overlap weight definition to better suit the data analytic goals.

Corrected overlap weight

For this we introduce a quantity

$$\mu = \max_{e \in \mathcal{E}} t(e)$$

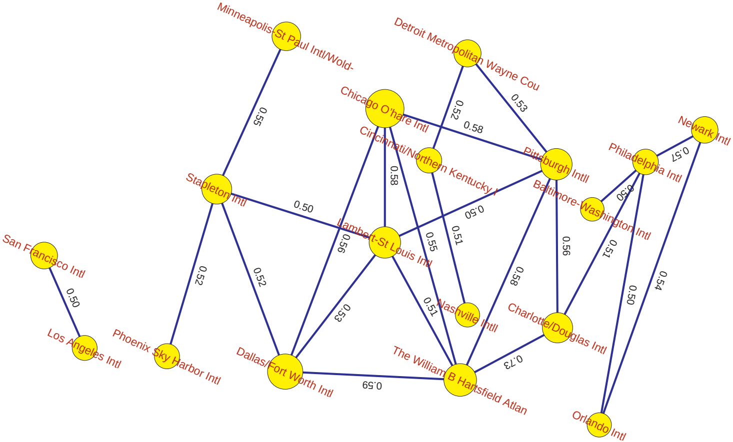
We define a *corrected overlap weight* as

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

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

US Airports links

with the largest corrected overlap weight, cut at 0.5



$\mu = 80$

US Airports links

with the largest corrected overlap weight

Corrected network measures

V. Batagelj

Introduction

Overlap weight

Corrected overlap weight

Clustering coefficient

Corrected clustering coefficient

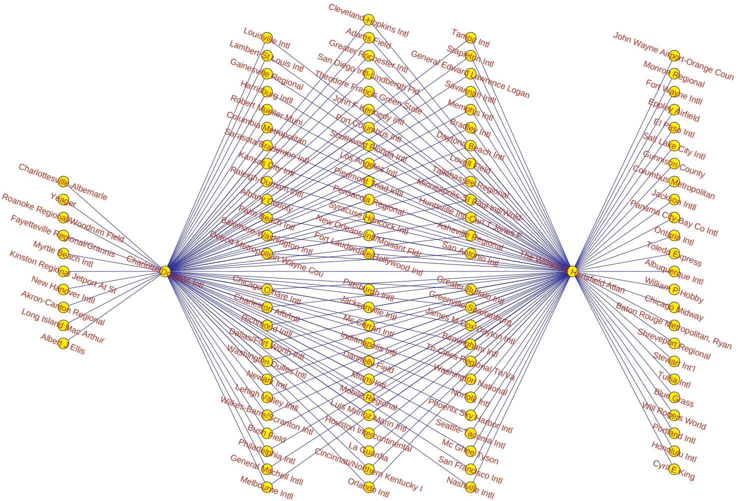
Conclusions

References

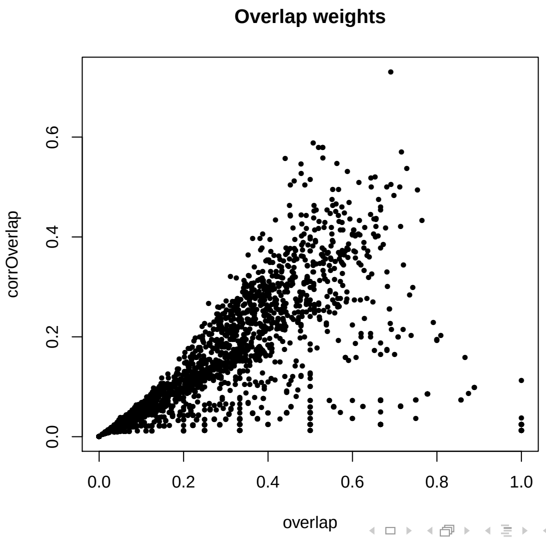
u	v		t(e)	d(u)	d(v)	o'(e)
The WB Hartsfield Atlan	Charlotte/Douglas Intl	=	76	101	87	0.73077
The WB Hartsfield Atlan	Dallas/Fort Worth Intl	=	73	101	118	0.58871
Chicago O'hare Intl	Pittsburgh Intll	=	80	139	94	0.57971
Chicago O'hare Intl	Lambert-St Louis Intl	=	80	139	94	0.57971
Dallas/Fort Worth Intl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

US Airports links

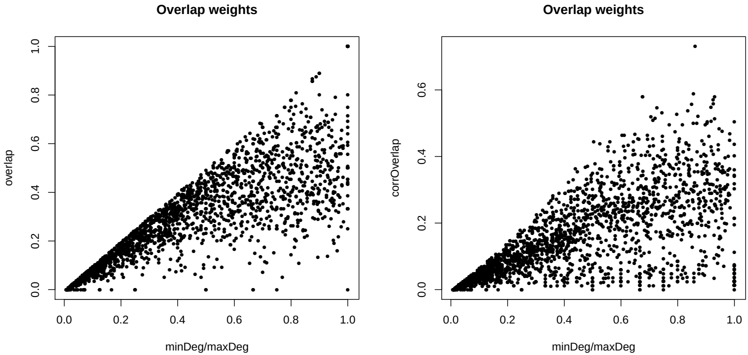
o'(WB Hartsfield Atlanta, Charlotte/Douglas Intl) = 0.7308



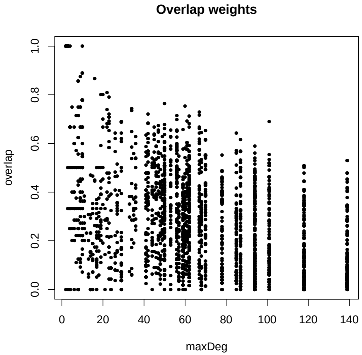
Comparison

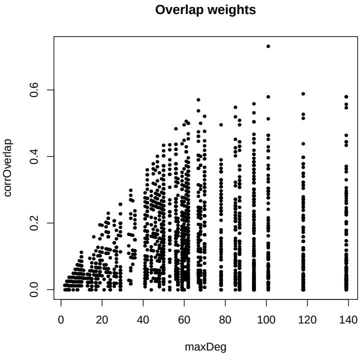

Overlap weights
corrOverlap
0.0
0.2
0.4
0.6
overlap
0.0
0.2
0.4
0.6
0.8
1.0

Comparison – $\text{minDeg}(e)/\text{maxDeg}(e)$

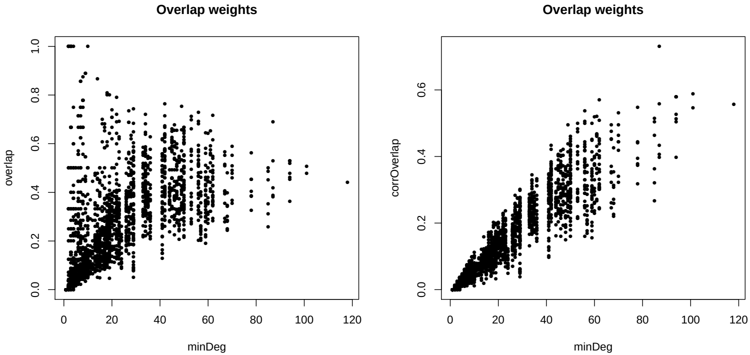


Comparison – maxDeg(e)





Comparison – minDeg(e)



Comparison – # of triangles

Corrected network measures

V. Batagelj

Introduction

Overlap weight

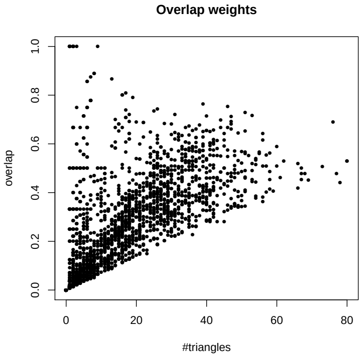
Corrected overlap weight

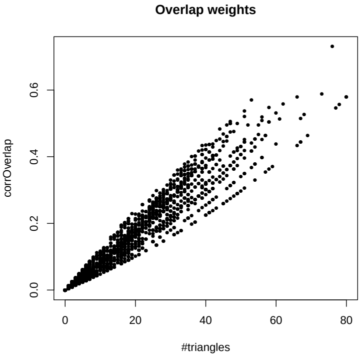
Clustering coefficient

Corrected clustering coefficient

Conclusions

References





Clustering coefficient

For a node $u \in \mathcal{V}$ in an undirected simple graph $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ its clustering coefficient is measuring a local density in node u and is defined as

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

where $N(u)$ is the set of neighbors of node 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 $S(u)$ 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 $K_{\deg(u)}$.

US Airports links with clustering coefficient = 1

1 Wiley Post-Will Rogers Mem
2 Ralph Wien Memorial
3 Aniak
4 Toledo Express
5 Myrtle Beach Intl
6 Rota Intl
7 Jack Mc Namara Field
8 Port Heiden
9 New Hanover Intll
10 Santa Maria Pub/Capt G Allan
11 Fayetteville Regional/Grannis
12 Lovell Field
13 St Paul Island
14 Elmira/Corning Regional
15 San Luis Obispo County-Mc Che
16 Binghamton Regional/Edwin A L
17 Fort Smith Regional
18 St Mary's
19 Asheville Regional
20 Molokai
21 Worcester Muni
22 Drake Field
23 Dubuque Regional
24 Tri-Cities Regional Tn/Va
25 Monterey Peninsula
26 Detroit City
27 Joplin Regional
28 Kwethluk
29 Hector Intll
30 Tompkins County
31 Cape Girardeau Regional
32 Merced Municipal/Macready Fie
33 King Salmon
34 Modesto City-County--Harry Sh
35 Natrona County Intl
36 Williamson County Regional
37 Deadhorse
38 Nome
39 Akiak
40 Dillingham
41 Evansville Regional
42 Charlottesville-Albemarle
43 Bishop Intll
44 Gunnison County
45 Friedman Memorial
46 Aspen-Pitkin Co/Sardy Field
47 Mbs Intll
48 Kwigillingok
49 Minot Intl
50 Pago Pago Intl
51 Babelthuap/Koror
52 Decatur
53 Quincy Muni Baldwin Field
54 Rafael Hernandez
55 Kongiganak
56 Bellingham Intl
57 La Crosse Muni
58 Hilo Intll
59 Rochester Intl
60 Kapalua
61 Lihue
62 Mc Allen Miller Intl
63 Rio Grande Valley Intl
64 Eareckson As
65 Corpus Christi Intl
66 St Petersburg/Clearwater In
67 Lehigh Valley Intll
68 Gainesville Regional
69 Burlington Regional
70 Lafayette Regional
71 Tuntutuliak
72 Tallahassee Regional
73 University Park
74 Sand Point
75 Tyler Pounds Field
76 Tweed-New Haven
77 Gregg County
78 Wilkes-Barre/Scranton Intl
79 Eastern Oregon Regional At
80 Stewart Intl

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 fast with increasing of $\deg(u)$.

Corrected clustering coefficient

To get a corrected version of the clustering coefficient we proposed in Pajek 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:

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

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

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

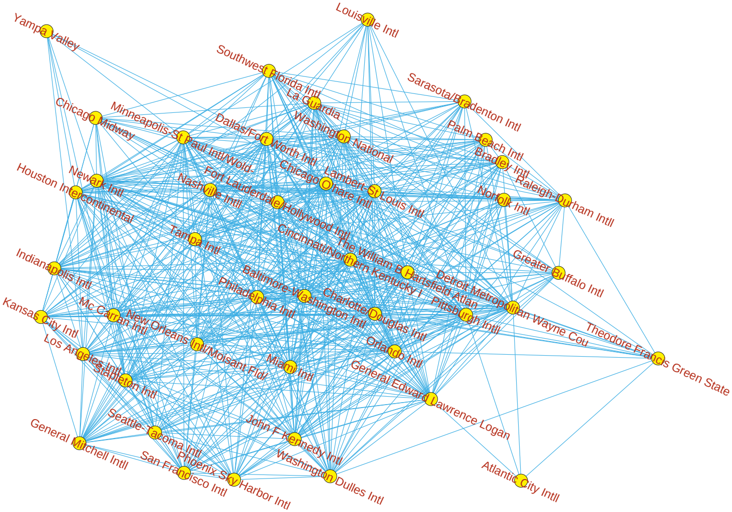
The value $cc'(u) = 1$ is attained in the case a on a μ-core, and in the case b on $K_{\mu+1}$.

US Airports links

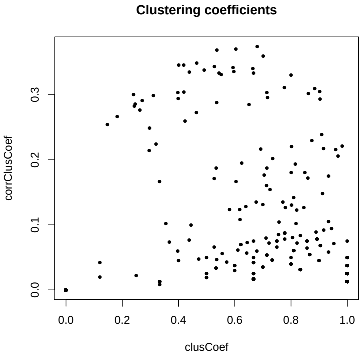
with the largest corrected clustering coefficient

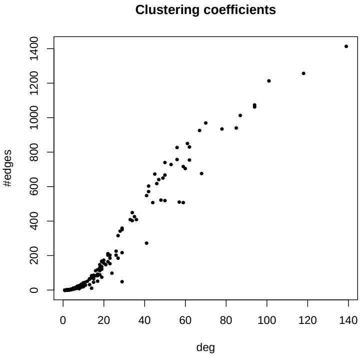
Rank	Value	Id	Rank	Value	Id
1	0.3739	Cleveland-Hopkins Intl	26	0.2990	Minneapolis-St Paul Intl/Wold-
2	0.3700	General Edward Lawrence Logan	27	0.2956	General Mitchell Intll
3	0.3688	Orlando Intl	28	0.2942	Phoenix Sky Harbor Intl
4	0.3595	Tampa Intl	29	0.2935	Palm Beach Intl
5	0.3488	Cincinnati/Northern Kentucky I	30	0.2914	Charlotte/Douglas Intl
6	0.3457	Detroit Metropolitan Wayne Cou	31	0.2881	Memphis Intl
7	0.3455	Newark Intl	32	0.2859	Lambert-St Louis Intl
8	0.3429	Baltimore-Washington Intl	33	0.2847	San Diego Intl-Lindbergh Fld
9	0.3415	Miami Intl	34	0.2824	Pittsburgh Intll
10	0.3405	Washington National	35	0.2762	Stapleton Intl
11	0.3379	Nashville Intll	36	0.2724	Washington Dulles Intl
12	0.3359	John F Kennedy Intl	37	0.2661	Dallas/Fort Worth Intl
13	0.3347	Philadelphia Intl	38	0.2595	Raleigh-Durham Intll
14	0.3335	Indianapolis Intl	39	0.2541	Chicago O'hare Intl
15	0.3335	La Guardia	40	0.2489	San Francisco Intl
16	0.3311	Mc Carran Intl	41	0.2386	Greater Buffalo Intl
17	0.3301	Fort Lauderdale/Hollywood Intl	42	0.2295	John Wayne Airport-Orange Coun
18	0.3106	New Orleans Intl/Moisant Fld/	43	0.2241	Seattle-Tacoma Intl
19	0.3095	Bradley Intl	44	0.2211	Sarasota/Bradenton Intl
20	0.3045	Port Columbus Intl	45	0.2207	Ontario Intl
21	0.3038	Los Angeles Intl	46	0.2175	Syracuse Hancock Intl
22	0.3036	Houston Intercontinental	47	0.2163	San Jose Intll
23	0.3036	Kansas City Intl	48	0.2158	Norfolk Intl
24	0.3017	Southwest Florida Intl	49	0.2144	Salt Lake City Intl
25	0.3002	The William B Hartsfield Atlan	50	0.2056	Greater Rochester Intl

Cleveland-Hopkins Intl neighbors

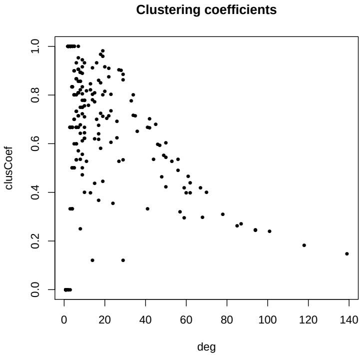


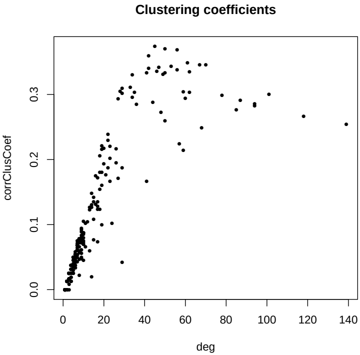
Comparison


Clustering coefficients
corrClusCoef
0.0
0.1
0.2
0.3
0.0
0.2
0.4
0.6
0.8
1.0
clusCoef


Clustering coefficients
#edges
0
200
400
600
800
1000
1200
1400
0
20
40
60
80
100
120
140
deg

Comparison – degrees





Conclusions

In the corrected measures we can replace μ with Δ. Its advantage is that it can be easier computed; but the corresponding measure is less ‘sensitive’.

References I

- P. W. Holland and S. Leinhardt (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.: Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences 104(18), 7332 (2007) paper
- D. J. Watts and Steven Strogatz (June 1998). "Collective dynamics of 'small-world' networks". Nature 393 (6684): 440–442.
- Wikipedia: Clustering coefficient
- Wikipedia: Overlap coefficient

Acknowledgments

This work was supported in part by the Slovenian Research Agency (research program P1-0294 and research projects J5-5537 and J1-5433).

The attendance on the CMStatistics (ERCIM) 2015 Conference was partially supported by the COST Action IC1408 – CRoNoS.