Corrected network elements importance measures

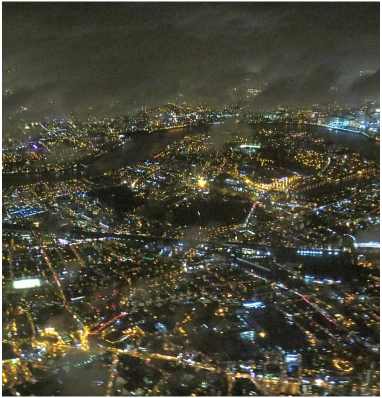
Vladimir Batagelj

IMFM Ljubljana and IAM UP Koper

1260. Sredin seminar

Ljubljana, 27. January 2016

Outline



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

Current version of slides (February 1, 2016, 05 : 33):
http://vlado.fmf.uni-lj.si/seminar/27jan16/sreda1260.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*

Examples: degrees, closeness, betweenness, hubs and authorities, clustering coefficient, etc.

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.

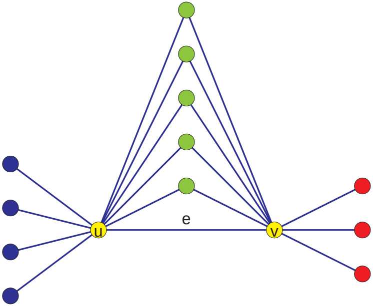
Overlap weight – triangles

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 their common neighbors (see figure). 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.



Overlap weight – definition

Two simple normalizations are:

$$\frac{t(e)}{n-2} \qquad \text{or} \qquad \frac{t(e)}{t_{max}}$$

where $n = |\mathcal{V}|$ is the number of nodes, and $t_{max} = \max_{e \in \mathcal{E}} t(e)$ 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.

Overlap weight – comments

The overlap weight is essentially a Jaccard similarity index

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

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. The operation $\oplus$ denotes the symmetric difference $X \oplus Y = (X \cup Y) \setminus (X \cap Y)$.

Overlap weight – comments

Another normalized overlap measure is the *overlap index*

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

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 end-node by the other.

Overlap weight – comments

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 $o(e) = 0$.

Overlap weight – properties

For every edge $e \in \mathcal{E}$ 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 $M(e) = t(e)$; and the value $o(e) = 0$ exactly when $t(e) = 0$.

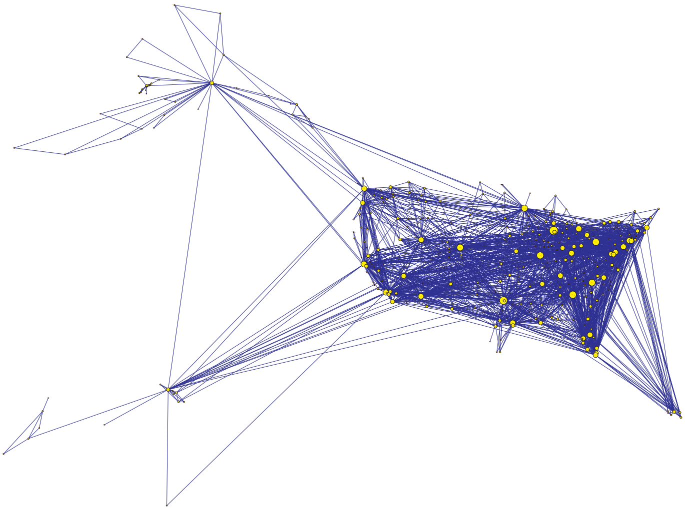
US Airports links 1997

Let us apply the overlap weight to the network of US Airports 1997 (see the figure). 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 the figure is proportional to its degree – the number of airports linked to it. The airports with the largest degree are:

deg(Chicago O'hare Intl) = 139,
deg(Dallas/Fort Worth Intl) = 118,
deg(The William B Hartsfield Atlanta) = 101,
deg(Lambert-St Louis Intl) = 94,
deg(Pittsburgh Intll) = 94, etc.

US Airports links 1997



US Airports links 1997

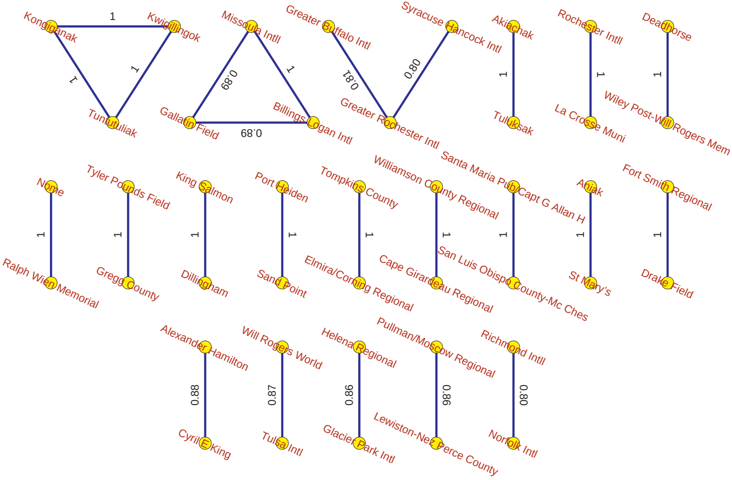
For the overlap weight the edge cut at level 0.8 (a subnetwork of all edges with overlap weight at least 0.8) is presented in the figure. It consists of two triangles, a 2-path, and 17 separate edges.

A tetrahedron (Kwigillingok, Kongiganak,Tuntutuliak, Bethel), see the figure and details in cut-out, gives the first triangle in the figure of cuts – attached with the node Bethel to the rest of network.

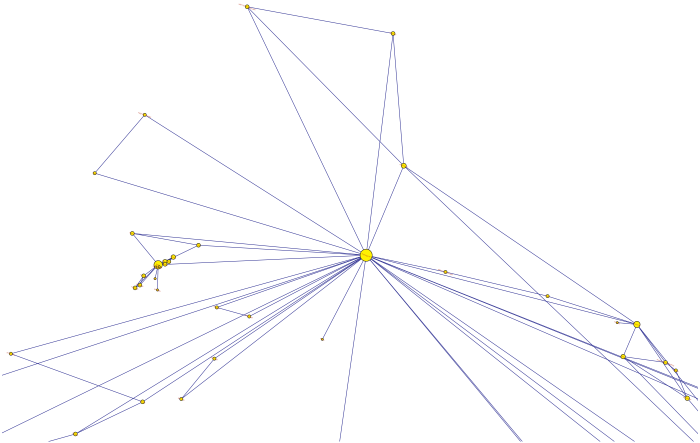
From this example we see that in real-life networks edges with the largest overlap weight tend to be edges with relatively small degrees in their end-nodes ($o(e) = 1$ implies $\deg(u) = \deg(v) = t(e) + 1$). Because of this the overlap weight is not very useful for data analytic tasks in searching for 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 only question its appropriateness for determining the most overlaped edges. We will try to improve the overlap weight definition to better suit the data analytic goals.

Edges with the largest overlap

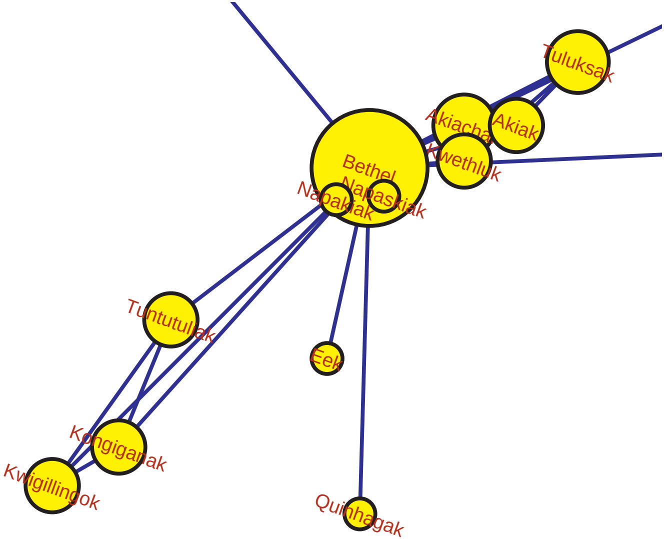
cut at 0.8



Zoom in



Zoom in


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

Corrected overlap weight

We introduce a quantity

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

and 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 $0 \leq o'(e) \leq 1$. $o'(e) = 0$ exactly when $t(e) = 0$, and $o'(e) = 1$ exactly when $\mu = M(e) = t(e)$.

US Airports links

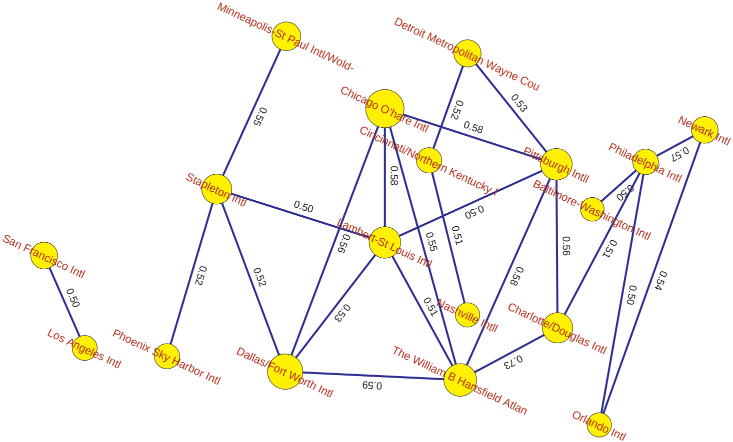
with the corrected overlap weight

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 the figure. Six links with the largest triangular weights are given in the table:

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

US Airports links

with the largest corrected overlap weight, cut at 0.5



$\mu = 80$

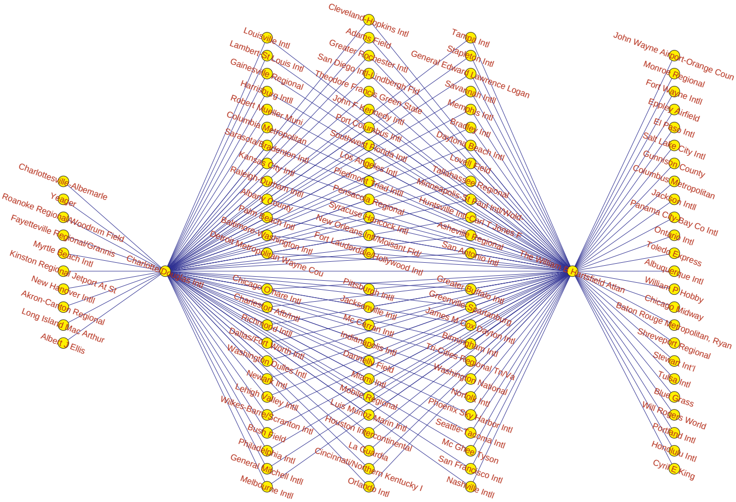
US Airports links

with the largest corrected overlap weight

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

US Airports links

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



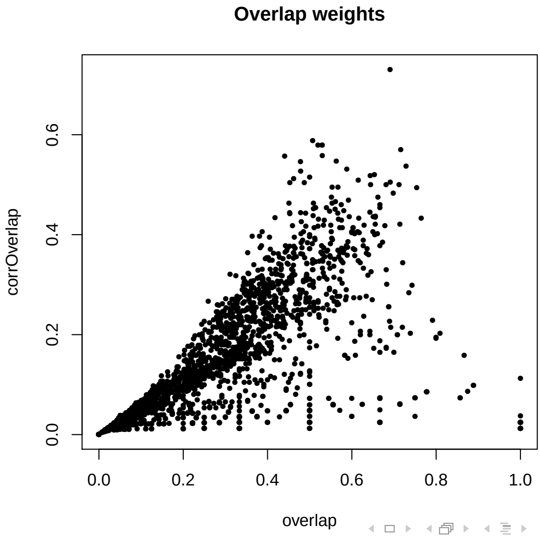
Comparisons

In the following figure 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 $o(e) < o'(e) \Leftrightarrow \mu < m(e)$.

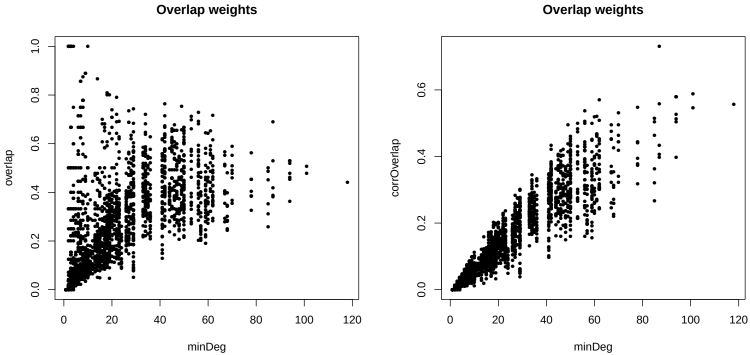
In the second figure the sets $\{(m(e), o(e)) : e \in \mathcal{E}\}$ and $\{(m(e), o'(e)) : e \in \mathcal{E}\}$ are displayed for the US Airports 1997 network. With increasing the $m(e)$ the corresponding overlap weight $o(e)$ is decreasing; and the corresponding corrected overlap weight $o'(e)$ is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

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

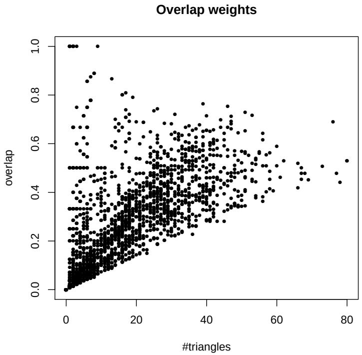
Comparison

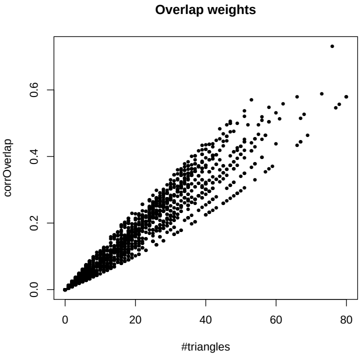

Overlap weights
corrOverlap
0.0
0.2
0.4
0.6
overlap
0.0
0.2
0.4
0.6
0.8
1.0

Comparison – minDeg(e)



Comparison – # of triangles





Clustering coefficient

For a node $u \in \mathcal{V}$ in an undirected simple graph $\mathbf{G} = (\mathcal{V}, \mathcal{E})$ its (local) clustering coefficient 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 $N(u)$ is the set of neighbors of node u and $E(u) = |\mathcal{E}(N(u))|$. 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 $S(u)$ 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.

US Airports links with clustering coefficient = 1

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

rank	deg	airport	rank	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

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

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

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

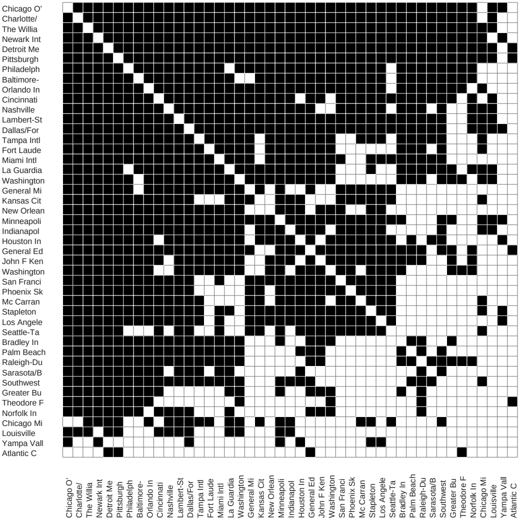
$$2 \cdot E(u) \leq \deg(u) \cdot (\deg(u) - 1) \leq \mu \cdot \deg(u)$$

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

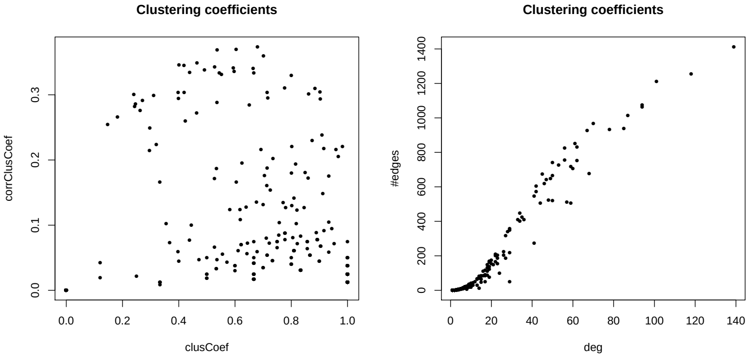
US Airports 1997 nodes
with the largest corrected clustering coefficient

Rank	Value	Id
1	0.3739	Cleveland-Hopkins Intl
2	0.3700	General Edward Lawrence Logan
3	0.3688	Orlando Intl
4	0.3595	Tampa Intl
5	0.3488	Cincinnati/Northern Kentucky I
6	0.3457	Detroit Metropolitan Wayne Cou
7	0.3455	Newark Intl
8	0.3429	Baltimore-Washington Intl
9	0.3415	Miami Intl
10	0.3405	Washington National
11	0.3379	Nashville Intll
12	0.3359	John F Kennedy Intl
13	0.3347	Philadelphia Intl
14	0.3335	Indianapolis Intl
15	0.3335	La Guardia

Cleveland-Hopkins Intl neighbors

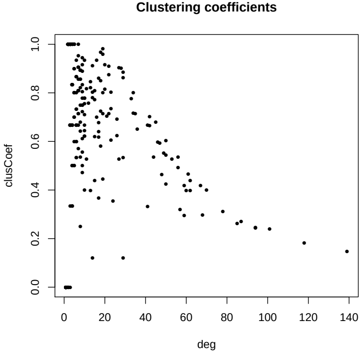


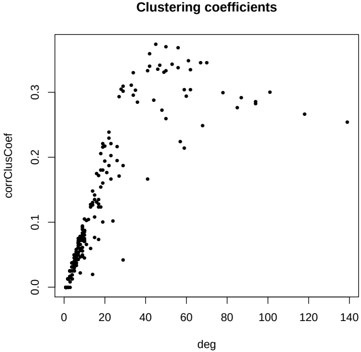
Comparison



In the left figure 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. In the right figure we also see that the number of edges in a node's neighborhood is almost functionally dependent on its degree.

Comparison – degrees





From the left figure we see that the clustering coefficient is decreasing with the increasing degree. Nodes with large degree have small values of clustering coefficient. The values of corrected clustering coefficient are large for nodes of large degree.

Widespread index

Motivation

SocNet: January 12, 2016

```
I need to have a measure of how widespread is the distribution of a
node attribute in a network. Let me explain:

My nodes have a textual attribute, let's say "preferred flavor for ice
cream".

I would like to know to what extent the flavor "raspberry" is a value
which is evenly distributed in the network, or to the contrary, just
found in one community. A low value would mean that only nodes
from a subregion of the network have this taste, a higher value
would show an even distribution of the value across the whole
network. I imagine that a difficulty is to account for the frequency
of the attribute: if many nodes of the network have "raspberry" for
the value of the attribute, it will tend to make this value distributed
more widely.

Any help or pointer on this would be very much appreciated!

Thank you, Clement Levallois
```

Widespread index

First attempt

Let us try to define a widespread index and show how it can be computed in Pajek.
Let S be a selected subset of set of nodes V in a simple (no parallel links) network $\mathcal{N} = (V, L)$. With $N(S)$ we denote the set of neighbors of the set S:

$$N(S) = \{u \in V : \exists v \in S : (v, u) \in L\}$$

and with $N^+(S) = S \cup N(S)$. With $n = |V|$ we denote the number of nodes.
A simple widespread index is

$$W_1(S) = \frac{|N^+(S)|}{n}.$$

Widespread index

Properties

We have:

- $0 \leq W_1(S) \leq 1$.
- $W_1(S) = 1$ iff S is a dominating set (Wikipedia). D is a dominating set of a network $N = (V, L)$ iff $N^+(D) = V$.
- if $S_1 \subset S_2$ then $N^+(S_1) \subseteq N^+(S_2)$.
- $|N^+(S_1)| \leq |N^+(S_2)|$ iff $W_1(S_1) \leq W_1(S_2)$.

Related to dominating sets is a domination number $\gamma(N)$

$$\gamma(N) = \ min_{D \text{ is a dominating set of} N} |D|$$

for which it holds $n \geq \gamma(N) \geq \lceil n/(1 + \Delta) \rceil \geq 1$, where Δ is the largest (out)degree in N.

Widespread index

Properties

Even better lower bound is $\gamma(N) \geq k$, where k is the smallest number such that

$$\sum_{i=1}^{k} (1 + d_i) \geq n$$

where (d_i) is a sequence of outdegrees ordered in decreasing order.

Note: In network with non-empty set of arcs the nodes with zero indegree are all in any dominant set. Let D_0 be the set of all such nodes. Then we get a better lower bound for γ – it is $|D_0| + k'$, where k' is the smallest number such that

$$\sum_{i=1}^{k'} (1 + d_i) \geq n - |N^+(D_0)|.$$

Since $V \setminus D_0$ can have zero degree nodes again we iterate the process. The final dis are computed in the final $V \setminus D_0$.

Widespread index

Dominant sets

The problem with the definition of $W_1(S)$ is that it doesn't consider the size of the set S. Let D^* be a minimal dominating set. Then an alternative definition of index could be

$$W^*(S) = \frac{|N(S) \setminus S|}{|V \setminus D^*|} = \frac{|N(S) \setminus S|}{n - \gamma}.$$

It is easy to see that

- $0 \leq W^*(S) \leq 1$.
- in a weakly connected network N : $W^*(S) = 1$ iff S is a minimal dominating set.
- if $|N^+(S_1)| = |N^+(S_2)|$ and $|S_1| < |S_2|$ then $W^*(S1) > W^*(S2)$.

Widespread index

Second attempt

Unfortunately the problem of determining the domination number $\gamma(N)$ is NP-complete – there is no efficient algorithm to compute W^*.

To get an efficiently computable index we could replace γ with 1 (it always holds $\gamma \geq 1$); or, even better, with k:

$$W_2(S) = \frac{|N(S) \setminus S|}{n-k}.$$

It holds:

- $W^*(S) \geq W_2(S)$.
- $0 \leq W_2(S) \leq 1$.
- $W_2(S) = 1$ iff S is a minimal dominant and independent (Wikipedia) set with k nodes.
- $W^*(S_1) > W^*(S_2)$ iff $W_2(S_1) > W_2(S_2)$.
- if $|N^+(S_1)| = |N^+(S_2)|$ and $|S_1| < |S_2|$ then $W_2(S_1) > W_2(S_2)$.

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.