Photo: Stefan Ernst,***Gartenkreuzspinne / Araneus diadematus***

Analysis of large networks
Islands

Vladimir Batagelj

University of Ljubljana

Slovenia

Dagstuhl Seminar 03361

Algorithmic Aspects of Large and Complex Networks

Dagstuhl, August 31 - September 5, 2003

Outline

Networks

A *network* $\mathbf{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ consists of:

- a *graph* $\mathbf{G} = (\mathcal{V}, \mathcal{L})$, where $\mathcal{V}$ is the set of *vertices* and $\mathcal{L}$ is the set of *lines* (links, ties). Undirected lines $\mathcal{E}$ are called *edges*, and directed lines $\mathcal{A}$ are called *arcs*. $n = \text{card}(\mathcal{V})$, $m = \text{card}(\mathcal{L})$.
- $\mathcal{P}$ *vertex value functions* or properties: $p : \mathcal{V} \rightarrow A$
- $\mathcal{W}$ *line value functions* or weights: $w : \mathcal{L} \rightarrow B$

Cuts

The *vertex-cut* of a network $\mathbf{N} = (\mathcal{V}, \mathcal{L}, p)$, $p : \mathcal{V} \rightarrow \mathbb{R}$, at selected level t is a subnetwork $\mathbf{N}(t) = (\mathcal{V}', \mathcal{L}(\mathcal{V}'), p)$, determined by the set

$$\mathcal{V}' = \{v \in \mathcal{V} : p(v) \geq t\}$$

and $\mathcal{L}(\mathcal{V}')$ is the set of lines from $\mathcal{L}$ that have both endpoints in $\mathcal{V}'$.

The *line-cut* of a network $\mathbf{N} = (\mathcal{V}, \mathcal{L}, w)$, $w : \mathcal{V} \rightarrow \mathbb{R}$, at selected level t is a subnetwork $\mathbf{N}(t) = (\mathcal{V}(\mathcal{L}'), \mathcal{L}', w)$, determined by the set

$$\mathcal{L}' = \{e \in \mathcal{L} : w(e) \geq t\}$$

and $\mathcal{V}(\mathcal{L}')$ is the set of all endpoints of the lines from $\mathcal{L}'$.

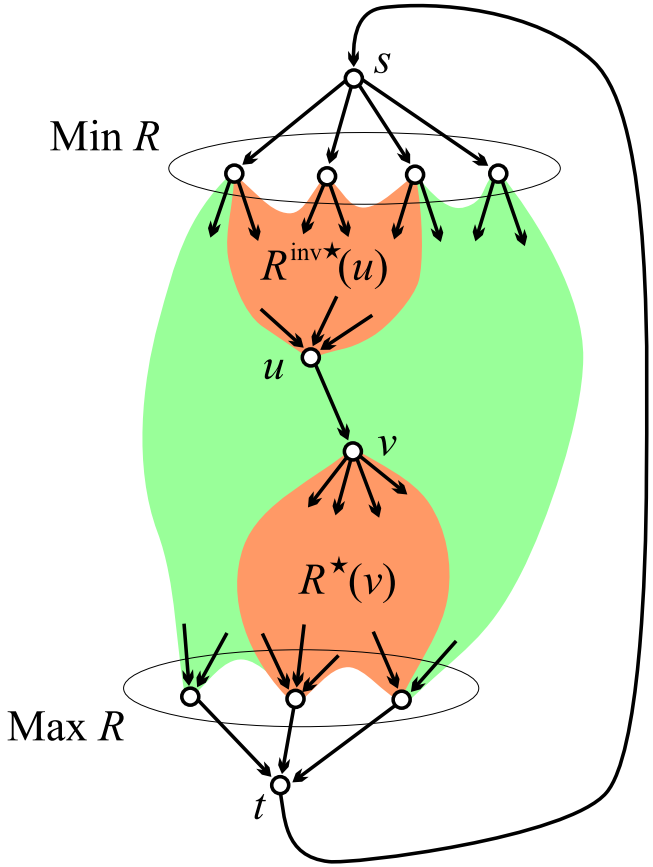
Simple analysis using cuts

We look at the components of $\mathbf{N}(t)$.

Their number and sizes depend on t. Usually there are many small components. Often we consider only components of size at least k and not exceeding K. The components of size smaller than k are discarded as 'noninteresting'; and the components of size larger than K are cut again at some higher level.

The values of thresholds t, k and K are determined by inspecting the distribution of vertex/arc-values and the distribution of component sizes and considering additional knowledge on the nature of network or goals of analysis.

Citation networks



In a given set of units $\mathbf{U}$ (articles, books, works, ...) we introduce a *citing* relation $R \subseteq \mathbf{U} \times \mathbf{U}$

$$uRv \equiv v \text{ cites } u$$

which determines a *citation network* $\mathbf{N} = (\mathbf{U}, R)$.

A citing relation is usually *irreflexive* and (almost) *acyclic*.

Citation networks characteristics

network	n	m	m_0	n_0	n_C	k_C	h	Δ_i	Δ_o	2	3	4
DNA	40	60	0	1	35	3	11	7	5	0	0	0
Coupling	223	657	1	5	218	1	16	19	134	0	0	0
Small world	396	1988	0	163	233	1	16	60	294	0	0	0
Small & Griffith	1059	4922	1	35	1024	1	28	89	232	2	0	0
Cocitation	1059	4929	1	35	1024	1	28	90	232	2	0	0
Scientometrics	3084	10416	1	355	2678	21	32	121	105	5	2	1
Kroto	3244	31950	1	0	3244	1	32	166	3243	6	0	0
SOM	4470	12731	2	698	3704	27	24	51	735	11	0	0
Zewail	6752	54253	1	101	6640	5	75	166	227	38	1	2
Lederberg	8843	41609	7	519	8212	35	63	135	1098	54	4	0
Desalination	8851	25751	7	1411	7143	115	27	73	137	12	0	1
US patents	3774768	16522438	1	0	3764117	3627	32	779	770	0	0	0

In the table: $n = |\mathbf{U}|$ is the number of vertices; $m = |R|$ is the number of arcs; m_0 is the number of loops; n_0 is the number of isolated vertices; n_C is the size of the largest weakly connected component; k_C is the number of nontrivial weakly connected components; h is the depth of network (length of the longest path); Δ_i and Δ_o are the maximum input and output degree. The last three columns contain the numbers of strongly connected components (cyclic parts) of size 2, 3 and 4.

Citation weights

An approach to the analysis of citation network is to determine for each unit / arc its *importance* or *weight*. These values are used afterward to determine the essential substructures in the network.

The citation network analysis started with the paper of Garfield et al. (1964) in which on the example of Asimov's history of DNA, it was shown that the analysis "*demonstrated a high degree of coincidence between an historian's account of events and the citational relationship between these events*".

Some methods of assigning weights $w : R \rightarrow \mathbb{R}_0^+$ to arcs were proposed by Hummon and Doreian (1989).

In 1991 we developed an efficient algorithm to compute these weights.

Citation weights algorithm

To compute the H&D's weights we introduce a related *search path count* (SPC) method for which the weights $N(u,v)$, uRv count the number of different paths from s to t through the arc (u,v).

To compute $N(u,v)$ we introduce two auxiliary quantities: let $N^-(v)$ denotes the number of different s-v paths, and $N^+(v)$ denotes the number of different v-t paths.

Every s-t path π containing the arc $(u,v) \in R$ can be uniquely expressed in the form

$$\pi = \sigma \circ (u,v) \circ \tau$$

where σ is a s-u path and τ is a v-t path. Since also every pair (σ,τ) of s-u / v-t paths gives a corresponding s-t path it follows:

$$N(u,v) = N^-(u) \cdot N^+(v), \qquad (u,v) \in R$$

Citation weights algorithm

where

$$N^-(u) = \begin{cases} 1 & u = s \\ \sum_{v:vRu} N^-(v) & \text{otherwise} \end{cases}$$

and

$$N^+(u) = \begin{cases} 1 & u = t \\ \sum_{v:uRv} N^+(v) & \text{otherwise} \end{cases}$$

This is the basis of an efficient algorithm for computing the weights $N(u, v)$ – after the topological sort of the network we can compute, using the above relations in topological order, the weights in time of order $O(m)$.

Using auxiliary counters we can also define a vertex-value function

$$p_c(v) = N^-(v) \cdot N^+(v)$$

as the number of paths going through the vertex v.

Properties of SPC weights

For the flow $N(u,v)$ the *Kirchoff's node law* holds:

For every node v in a citation network in standard form it holds

$$\text{incoming flow} = \text{outgoing flow} = p_c(v)$$

The *total flow* through the citation network equals $N(t,s)$. This gives us a natural way to normalize the weights

$$w(u,v) = \frac{N(u,v)}{N(t,s)} \quad \Rightarrow \quad 0 \leq w(u,v) \leq 1$$

If C is a minimal arc-cut-set $\sum_{(u,v)\in C} w(u,v) = 1$.

Very large/small numbers that result as the SPC weights in large networks are not easy to use. One possibility to overcome this problem is to use the logarithms of the obtained weights.

Example – SOM

As an example we shall analyze the SOM (self-organizing maps) literature network obtained from Garfield's collection of citation networks.

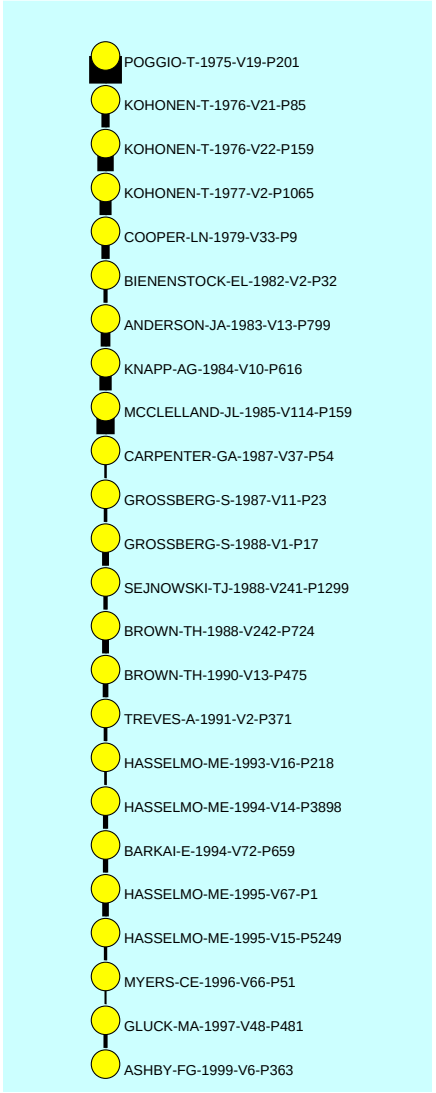
The analysis was done with program Pajek.

We read the citation network (with additional information) `Kohonen.paj`. The network has 4470 vertices and 12731 arcs (2 loops). First we test the network for acyclicity. Since there are 11 nontrivial strong components we eliminate them by shrinking each component into a single vertex. This operation produces some loops that should be removed.

Now we can compute the SPC citation weights. Pajek returns the following results: the network with citation weights on arcs, the main path network and the vector with vertex values.

First we draw the main path network using macro `Layers`. We compute also the CPM path and draw it.

SOM citation main path and CPM path



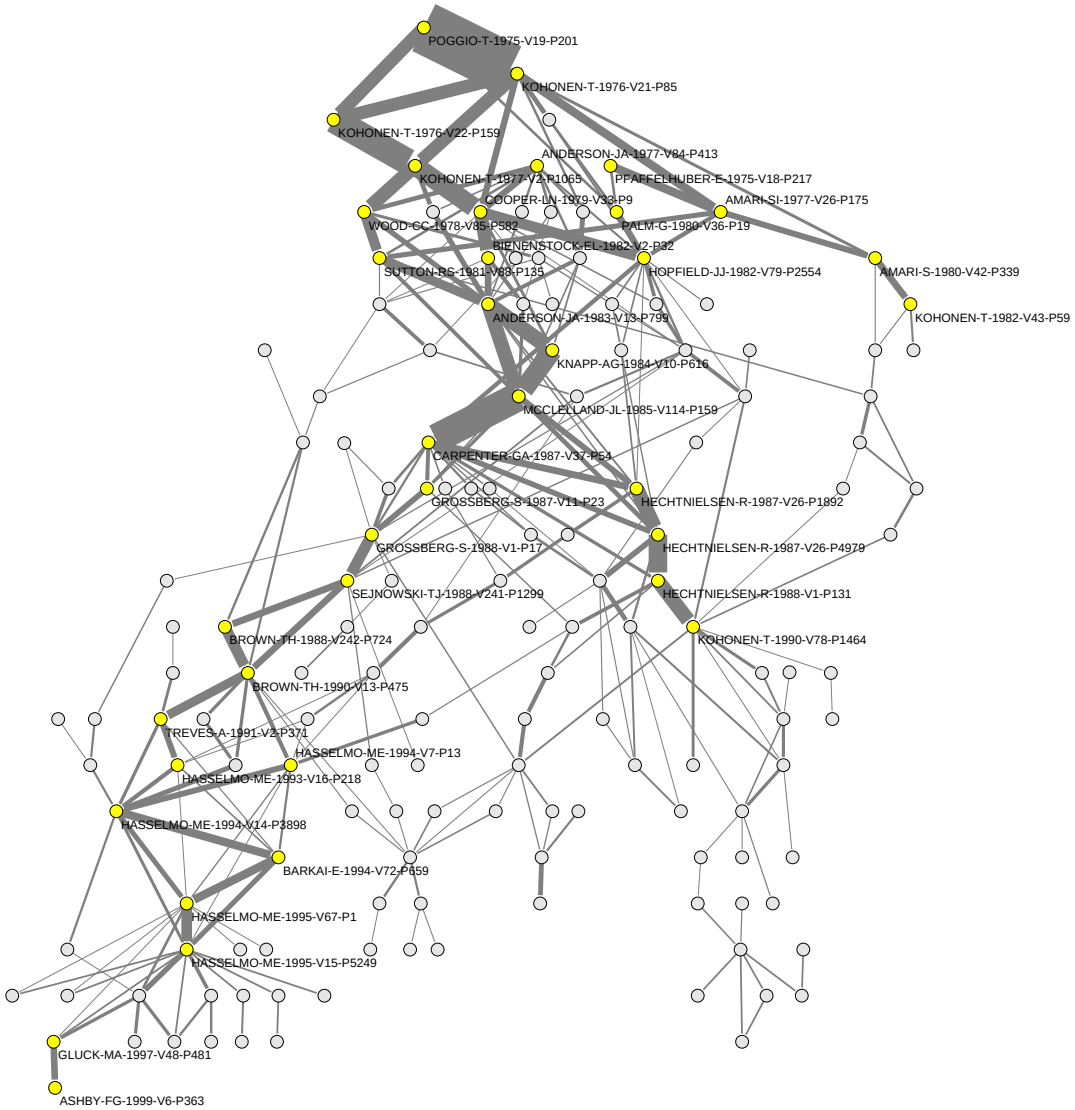
SOM main subnetwork

Inspecting the distribution of values of weights on arcs (lines) we select a threshold 0.007 and delete all arcs with weights lower than selected threshold. We delete also all isolated vertices (degree $= 0$) and small ($k = 5$) components. A single component remains. We draw it. We improve the obtained layout manually.

We label only the 'important' vertices – endpoints of arcs with weight at least 0.05.

From the picture we see that there isn't a single stream in the development of SOM field.

SOM arc-cut subnetwork at level 0.007



Islands (with M. Zaveršnik)

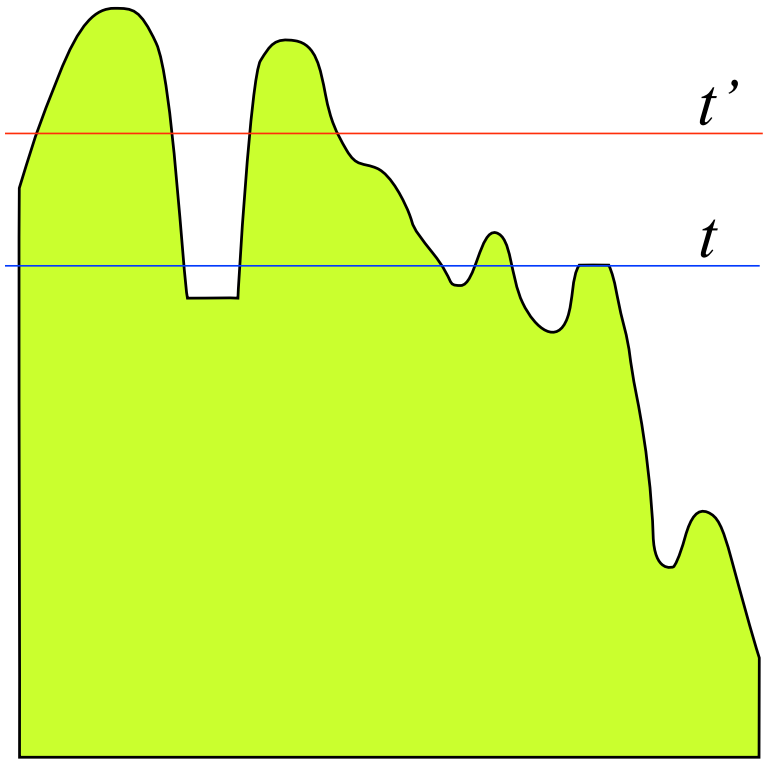
Using the line weights we can define an *island* as a connected small subnetwork of size in the interval k .. K with stronger internal cohesion relatively to its neighborhood.

We can base an islands search procedure on cuts. In the reduced network, for a selected threshold t, we determine (weakly) connected components. Each component of size in range $k..K$ represents an island since:

- they are connected and of selected size,
- all lines linking them to their outside neighbors have weight lower than t, and
- each vertex of an island is linked with some other vertex in the same island with a line with a weight at least t.

In similar way vertex value islands can be defined.

Islands algorithm



We developed an algorithm that identifies all maximal (k, K)-islands in a given network; and extended it for all single peak islands.

Each island is identified with its *port* – its lowest vertex.

The main problem are the vertices at the same level – *flat* regions.

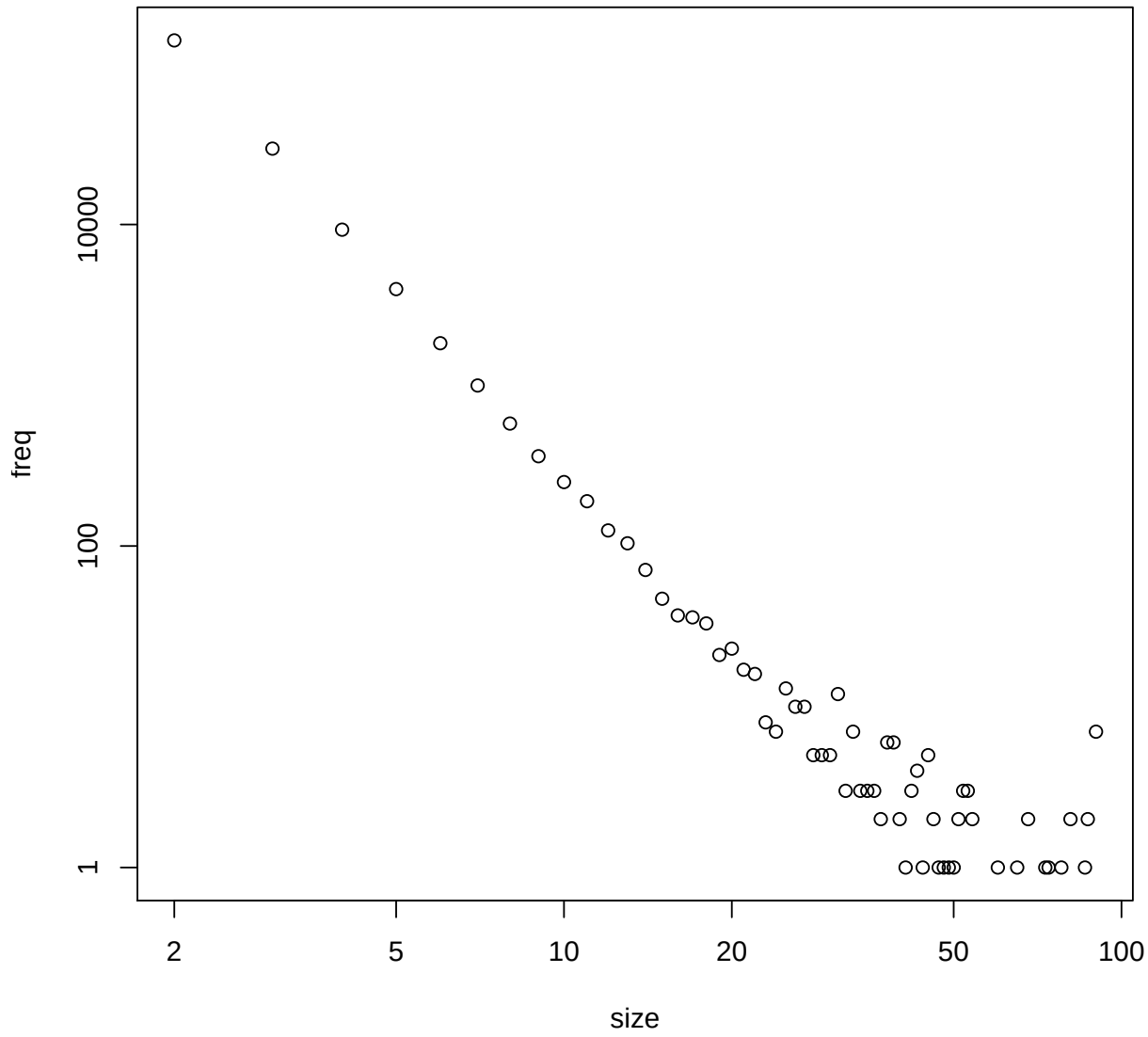
Example – US patents

As an example, let us look at Nber network of US Patents. It has 3774768 vertices and 16522438 arcs (1 loop). We computed SPC weights in it and determined all (2,90)-islands. The reduced network has 470137 vertices, 307472 arcs and for different k: C_2 =187610, C_5 =8859,C_{30} =101, C_{50} =30 islands. Rolex

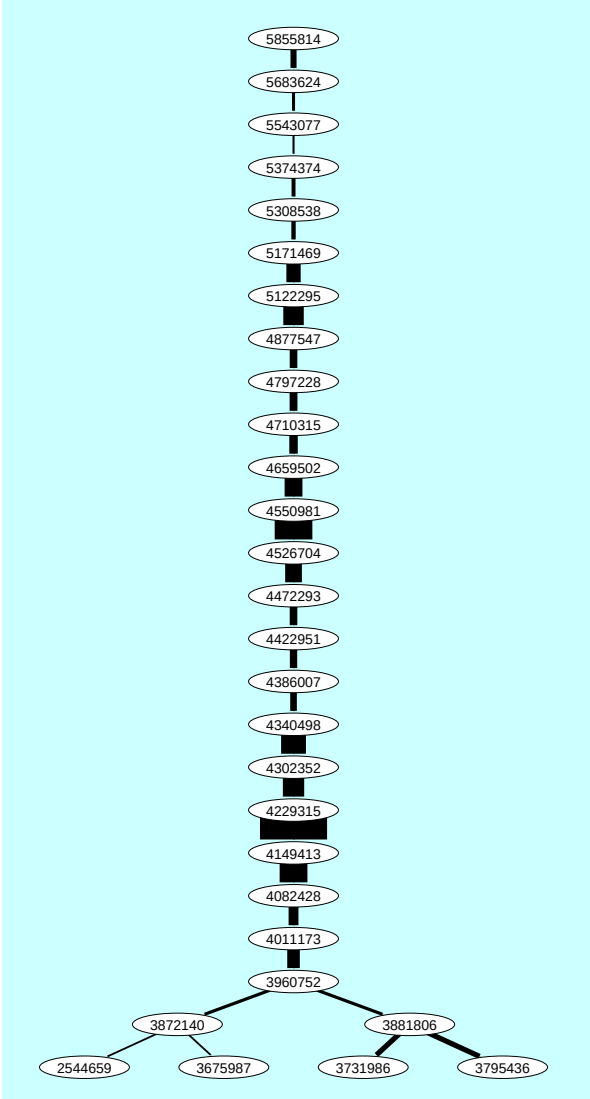
```
 [1]   0 139793  29670 9288 3966 1827 997 578 362 250
[11] 190    125    104   71   47   37  36  33  21  23
[21]  17     16      8    7   13   10  10   5   5   5
[31]  12      3      7    3    3    3   2   6   6   2
[41]   1      3      4    1    5    2   1   1   1   1
[51]   2      3      3    2    0    0   0   0   0   1
[61]   0      0      0    0    1    0   0   2   0   0
[71]   0      0      1    1    0    0   0   1   0   0
[81]   2      0      0    0    0    1   2   0   0   7
```

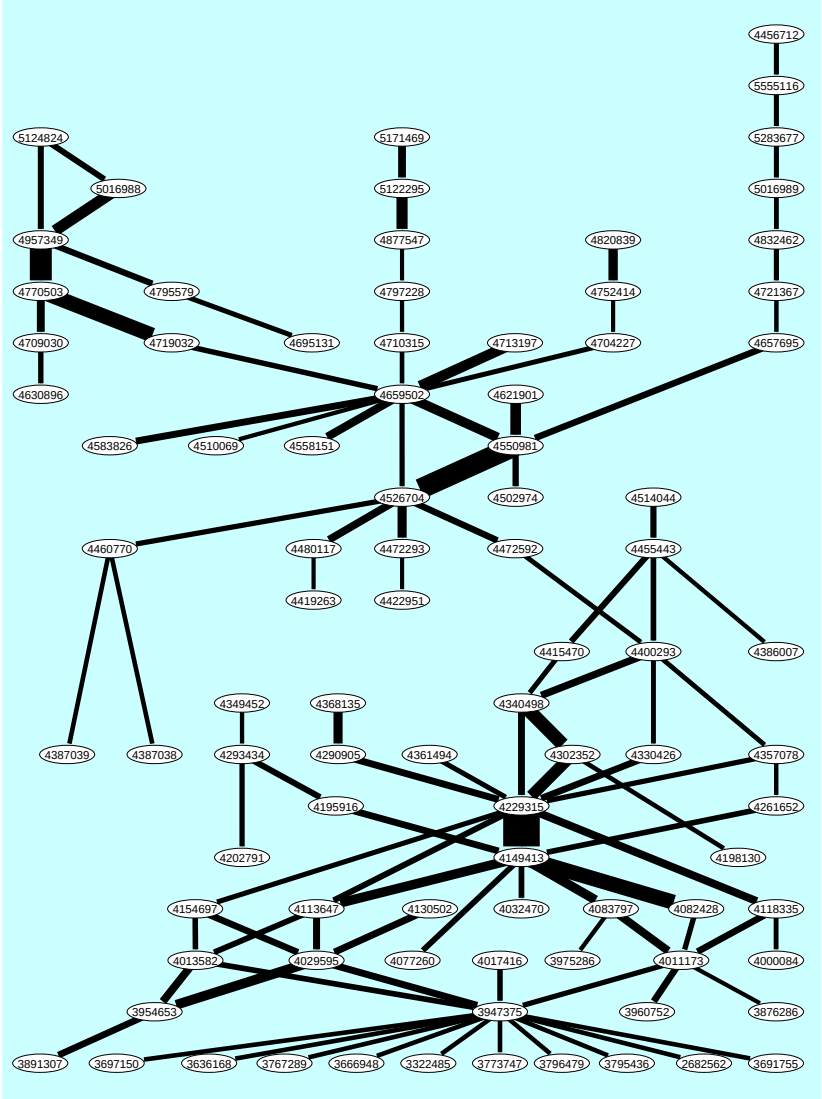
```
t <- scan("spcgs.clu",skip=1)
f <- tabulate(tabulate(t))
c <- f[f>0]
i <- (1:length(f))[f>0]
plot(i,c,log='xy',main='Island size distribution',
     xlab='size',ylab='freq')
```

Island size distribution



Main path and main island of Patents





Liquid crystal display

Table 1: Patents on the liquid-crystal display

patent	date	author(s) and title
2544659	Mar 13, 1951	Dreyer. Dichroic light-polarizing sheet and the like and the formation and use thereof
2682562	Jun 29, 1954	Wender, et al. Reduction of aromatic carbinols
3322485	May 30, 1967	Williams. Electro-optical elements utilazing an organic nematic compound
3636168	Jan 18, 1972	Josephson. Preparation of polynuclear aromatic compounds
3666948	May 30, 1972	Mechlowitz, et al. Liquid crystal termal imaging system having an undisturbed image on a disturbed background
3675987	Jul 11, 1972	Rafuse. Liquid crystal compositions and devices
3691755	Sep 19, 1972	Girard. Clock with digital display
3697150	Oct 10, 1972	Wysochi. Electro-optic systems in which an electrophoretic-like or dipolar material is dispersed throughout a liquid crystal to reduce the turn-off time
3731986	May 8, 1973	Fergason. Display devices utilizing liquid crystal light modulation
3767289	Oct 23, 1973	Aviram, et al. Class of stable trans-stilbene compounds, some displaying nematic mesophases at or near room temperature and others in a range up to 100°C
3773747	Nov 20, 1973	Steinstrasser. Substituted azoxy benzene compounds
3795436	Mar 5, 1974	Boller, et al. Nematogenic material which exhibit the Kerr effect at isotropic temperatures
3796479	Mar 12, 1974	Helfrich, et al. Electro-optical light-modulation cell utilizing a nematogenic material which exhibits the Kerr effect at isotropic temperatures
3872140	Mar 18, 1975	Klanderman, et al. Liquid crystalline compositions and method
3876286	Apr 8, 1975	Deutscher, et al. Use of nematic liquid crystalline substances
3881806	May 6, 1975	Suzuki. Electro-optical display device
3891307	Jun 24, 1975	Tsukamoto, et al. Phase control of the voltages applied to opposite electrodes for a cholesteric to nematic phase transition display
3947375	Mar 30, 1976	Gray, et al. Liquid crystal materials and devices
3954653	May 4, 1976	Yamazaki. Liquid crystal composition having high dielectric anisotropy and display device incorporating same
3960752	Jun 1, 1976	Klanderman, et al. Liquid crystal compositions
3975286	Aug 17, 1976	Oh. Low voltage actuated field effect liquid crystals compositions and method of synthesis
4000084	Dec 28, 1976	Hsieh, et al. Liquid crystal mixtures for electro-optical display devices
4011173	Mar 8, 1977	Steinstrasser. Modified nematic mixtures with positive dielectric anisotropy
4013582	Mar 22, 1977	Gavrilovic. Liquid crystal compounds and electro-optic devices incorporating them
4017416	Apr 12, 1977	Inukai, et al. P-cyanophenyl 4-alkyl-4'-biphenylcarboxylate, method for preparing same and liquid crystal compositions using same
4029595	Jun 14, 1977	Ross, et al. Novel liquid crystal compounds and electro-optic devices incorporating them
4032470	Jun 28, 1977	Bloom, et al. Electro-optic device
4077260	Mar 7, 1978	Gray, et al. Optically active cyano-biphenyl compounds and liquid crystal materials containing them
4082428	Apr 4, 1978	Hsu. Liquid crystal composition and method

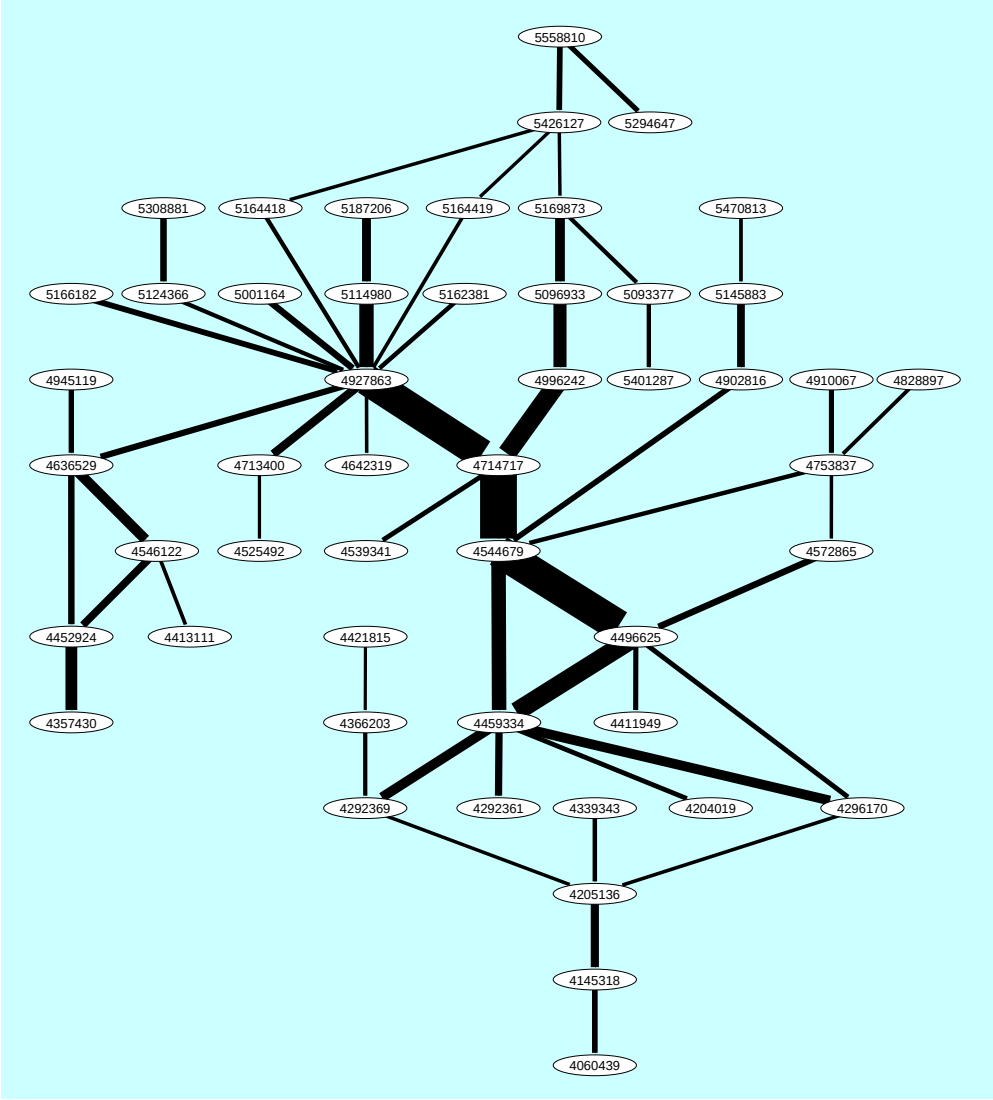
Table 2: Patents on the liquid-crystal display

patent	date	author(s) and title
4083797	Apr 11, 1978	Oh. Nematic liquid crystal compositions
4113647	Sep 12, 1978	Coates, et al. Liquid crystalline materials
4118335	Oct 3, 1978	Krause, et al. Liquid crystalline materials of reduced viscosity
4130502	Dec 19, 1978	Eidenschink, et al. Liquid crystalline cyclohexane derivatives
4149413	Apr 17, 1979	Gray, et al. Optically active liquid crystal mixtures and liquid crystal devices containing them
4154697	May 15, 1979	Eidenschink, et al. Liquid crystalline hexahydroterphenyl derivatives
4195916	Apr 1, 1980	Coates, et al. Liquid crystal compounds
4198130	Apr 15, 1980	Boller, et al. Liquid crystal mixtures
4202791	May 13, 1980	Sato, et al. Nematic liquid crystalline materials
4229315	Oct 21, 1980	Krause, et al. Liquid crystalline cyclohexane derivatives
4261652	Apr 14, 1981	Gray, et al. Liquid crystal compounds and materials and devices containing them
4290905	Sep 22, 1981	Kanbe. Ester compound
4293434	Oct 6, 1981	Deutscher, et al. Liquid crystal compounds
4302352	Nov 24, 1981	Eidenschink, et al. Fluorophenylcyclohexanes, the preparation thereof and their use as components of liquid crystal dielectrics
4330426	May 18, 1982	Eidenschink, et al. Cyclohexylbiphenyls, their preparation and use in dielectrics and electrooptical display elements
4340498	Jul 20, 1982	Sugimori. Halogenated ester derivatives
4349452	Sep 14, 1982	Osman, et al. Cyclohexylcyclohexanoates
4357078	Nov 2, 1982	Carr, et al. Liquid crystal compounds containing an alicyclic ring and exhibiting a low dielectric anisotropy and liquid crystal materials and devices incorporating such compounds
4361494	Nov 30, 1982	Osman, et al. Anisotropic cyclohexyl cyclohexylmethyl ethers
4368135	Jan 11, 1983	Osman. Anisotropic compounds with negative or positive DC-anisotropy and low optical anisotropy
4386007	May 31, 1983	Krause, et al. Liquid crystalline naphthalene derivatives
4387038	Jun 7, 1983	Fukui, et al. 4-(Trans-4'-alkylcyclohexyl) benzoic acid 4"'-cyano-4"-biphenylyl esters
4387039	Jun 7, 1983	Sugimori, et al. Trans-4-(trans-4'-alkylcyclohexyl)-cyclohexane carboxylic acid 4"'-cyanobiphenyl ester
4400293	Aug 23, 1983	Romer, et al. Liquid crystalline cyclohexylphenyl derivatives
4415470	Nov 15, 1983	Eidenschink, et al. Liquid crystalline fluorine-containing cyclohexylbiphenyls and dielectrics and electro-optical display elements based thereon
4419263	Dec 6, 1983	Praefcke, et al. Liquid crystalline cyclohexylcarbonitrile derivatives
4422951	Dec 27, 1983	Sugimori, et al. Liquid crystal benzene derivatives
4455443	Jun 19, 1984	Takatsu, et al. Nematic halogen Compound
4456712	Jun 26, 1984	Christie, et al. Bismaleimide triazine composition
4460770	Jul 17, 1984	Petrzilka, et al. Liquid crystal mixture
4472293	Sep 18, 1984	Sugimori, et al. High temperature liquid crystal substances of four rings and liquid crystal compositions containing the same
4472592	Sep 18, 1984	Takatsu, et al. Nematic liquid crystalline compounds
4480117	Oct 30, 1984	Takatsu, et al. Nematic liquid crystalline compounds
4502974	Mar 5, 1985	Sugimori, et al. High temperature liquid-crystalline ester compounds
4510069	Apr 9, 1985	Eidenschink, et al. Cyclohexane derivatives

Table 3: Patents on the liquid-crystal display

patent	date	author(s) and title
4514044	Apr 30, 1985	Gunjima, et al. 1-(Trans-4-alkylcyclohexyl)-2-(trans-4'-(p-substituted phenyl) cyclohexyl)ethane and liquid crystal mixture
4526704	Jul 2, 1985	Petrzilka, et al. Multiring liquid crystal esters
4550981	Nov 5, 1985	Petrzilka, et al. Liquid crystalline esters and mixtures
4558151	Dec 10, 1985	Takatsu, et al. Nematic liquid crystalline compounds
4583826	Apr 22, 1986	Petrzilka, et al. Phenylethanes
4621901	Nov 11, 1986	Petrzilka, et al. Novel liquid crystal mixtures
4630896	Dec 23, 1986	Petrzilka, et al. Benzonitriles
4657695	Apr 14, 1987	Saito, et al. Substituted pyridazines
4659502	Apr 21, 1987	Fearon, et al. Ethane derivatives
4695131	Sep 22, 1987	Balkwill, et al. Disubstituted ethanes and their use in liquid crystal materials and devices
4704227	Nov 3, 1987	Krause, et al. Liquid crystal compounds
4709030	Nov 24, 1987	Petrzilka, et al. Novel liquid crystal mixtures
4710315	Dec 1, 1987	Schad, et al. Anisotropic compounds and liquid crystal mixtures therewith
4713197	Dec 15, 1987	Eidenschink, et al. Nitrogen-containing heterocyclic compounds
4719032	Jan 12, 1988	Wachtler, et al. Cyclohexane derivatives
4721367	Jan 26, 1988	Yoshinaga, et al. Liquid crystal device
4752414	Jun 21, 1988	Eidenschink, et al. Nitrogen-containing heterocyclic compounds
4770503	Sep 13, 1988	Buchecker, et al. Liquid crystalline compounds
4795579	Jan 3, 1989	Vauchier, et al. 2,2'-difluoro-4-alkoxy-4'-hydroxydiphenyls and their derivatives, their production process and their use in liquid crystal display devices
4797228	Jan 10, 1989	Goto, et al. Cyclohexane derivative and liquid crystal composition containing same
4820839	Apr 11, 1989	Krause, et al. Nitrogen-containing heterocyclic esters
4832462	May 23, 1989	Clark, et al. Liquid crystal devices
4877547	Oct 31, 1989	Weber, et al. Liquid crystal display element
4957349	Sep 18, 1990	Clerc, et al. Active matrix screen for the color display of television pictures, control system and process for producing said screen
5016988	May 21, 1991	Iimura. Liquid crystal display device with a birefringent compensator
5016989	May 21, 1991	Okada. Liquid crystal element with improved contrast and brightness
5122295	Jun 16, 1992	Weber, et al. Matrix liquid crystal display
5124824	Jun 23, 1992	Kozaki, et al. Liquid crystal display device comprising a retardation compensation layer having a maximum principal refractive index in the thickness direction
5171469	Dec 15, 1992	Hittich, et al. Liquid-crystal matrix display
5283677	Feb 1, 1994	Sagawa, et al. Liquid crystal display with ground regions between terminal groups
5308538	May 3, 1994	Weber, et al. Supertwist liquid-crystal display
5374374	Dec 20, 1994	Weber, et al. Supertwist liquid-crystal display
5543077	Aug 6, 1996	Rieger, et al. Nematic liquid-crystal composition
5555116	Sep 10, 1996	Ishikawa, et al. Liquid crystal display having adjacent electrode terminals set equal in length
5683624	Nov 4, 1997	Sekiguchi, et al. Liquid crystal composition
5855814	Jan 5, 1999	Matsui, et al. Liquid crystal compositions and liquid crystal display elements

Producing a foam



Producing a foam

patent	date	author(s) and title
4060439	Nov 29, 1977	Rosemund, et al. Polyurethane foam composition and method of making same
4292369	Sep 29, 1981	Ohashi, et al. Fireproof laminates
4357430	Nov 2, 1982	VanCleve. Polymer/polyols, methods for making same and polyurethanes based thereon
4459334	Jul 10, 1984	Blanpied, et al. Composite building panel
4496625	Jan 29, 1985	Snider , et al. Alkoxylated aromatic amine-aromatic polyester polyol blend and polyisocyanurate foam therefrom
4544679	Oct 1, 1985	Tideswell, et al. Polyol blend and polyisocyanurate foam produced therefrom
4714717	Dec 22, 1987	Londrigan, et al. Polyester polyols modified by low molecular weight glycols and cellular foams therefrom
4927863	May 22, 1990	Bartlett, et al. Process for producing closed-cell polyurethane foam compositions expanded with mixtures of blowing agents
4996242	Feb 26, 1991	Lin. Polyurethane foams manufactured with mixed gas/liquid blowing agents
5169873	Dec 8, 1992	Behme, et al. Process for the manufacture of foams with the aid of blowing agents containing fluoroalkanes and fluorinated ethers, and foams obtained by this process
5308881	May 3, 1994	Londrigan, et al. Surfactant for polyisocyanurate foams made with alternative blowing agents
5558810	Sep 24, 1996	Minor, et al. Pentafluoropropane compositions

Conclusions

We proposed an approach to the analysis of networks that can be used also for very large networks with millions of vertices and arcs. The islands can be used as a *filter* to identify interesting subnetworks that are further analyzed using more sophisticated tools.

Analyses with other weights/values functions – for example:

$p =$ betweeness,

$p =$ core number,

$w =$ number of triangles containing the link, ...

Sources

Vladimir Batagelj, Andrej Mrvar: Pajek.
http://vlado.fmf.uni-lj.si/pub/networks/pajek/

Vladimir Batagelj: Efficient Algorithms for Citation Network Analysis.
http://arxiv.org/abs/cs.DL/0309023

Vladimir Batagelj: Papers on network analysis.
http://vlado.fmf.uni-lj.si/pub/networks/doc/

Pajek's datasets – citation networks:
http://vlado.fmf.uni-lj.si/pub/networks/data/cite/