



Photo: V. Batagelj, *Net*

Islands

Vladimir Batagelj
Matjaž Zaveršnik

University of Ljubljana

COSIN Meeting at the University of Karlsruhe

November 8th and 9th 2004

Outline

1	Networks	1
2	Cuts	2
4	Cuts and islands	4
5	Vertex islands	5
10	Simple vertex islands	10
12	Line islands	12
17	Simple line islands	17
19	Example: Reuters Terror News	19
20	Example: US Patents	20
21	Example: The Edinburgh Associative Thesaurus	21
27	Example: Amazon CDs and books networks	27
33	Conclusions	33

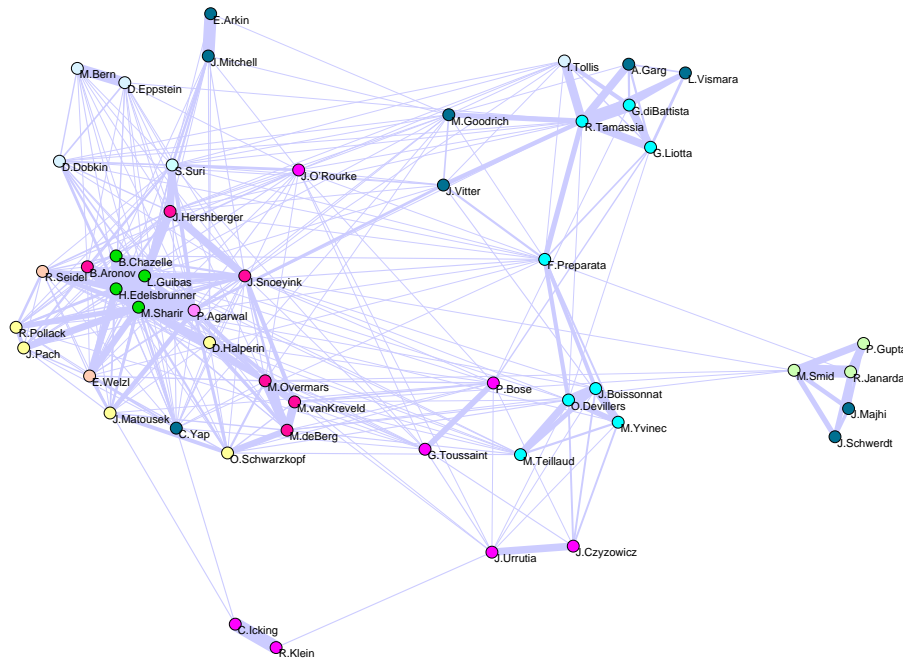
Networks

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

- a *graph* $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, where \mathcal{V} is the set of vertices and $\mathcal{L} = \mathcal{E} \cup \mathcal{A}$ 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* / properties: $p : \mathcal{V} \rightarrow A$
- \mathcal{W} *line value functions* / weights: $w : \mathcal{L} \rightarrow B$



Cuts

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

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

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 $\mathcal{N} = (\mathcal{V}, \mathcal{L}, w)$, $w: \mathcal{L} \rightarrow \mathbb{R}$, at selected level t is a subnetwork $\mathcal{N}(t) = (\mathcal{V}(\mathcal{L}'), \mathcal{L}', w)$, determined by

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

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

- The line-cut at level t is a vertex-cut at the same level for

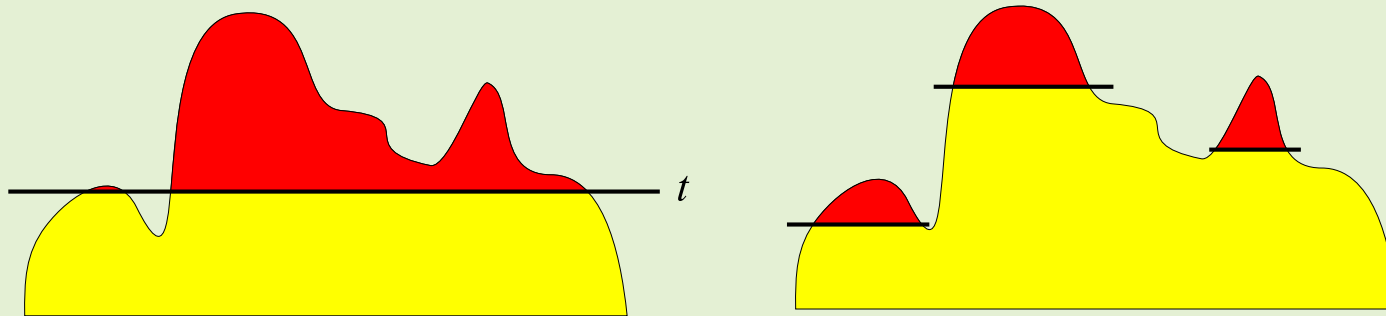
$$p(v) = \max_{u \in N(v)} w(v, u)$$

where we preserve only lines with $w(e) \geq t$.

Simple analysis using cuts

- After making a cut at selected level t we look at the components of the $\mathcal{N}(t)$. Their number and sizes depend on t . Usually there are many small and some large 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/line values and the distribution of component sizes and considering additional knowledge about the nature of network or goals of analysis.

Cuts and islands



Vertex islands

- Nonempty subset of vertices $\mathcal{C} \subseteq \mathcal{V}$ is a *vertex island*, if
 - the corresponding induced subgraph $\mathcal{G}|_{\mathcal{C}} = (\mathcal{C}, \mathcal{L}(\mathcal{C}))$ is connected, and
 - the values of the vertices in the neighborhood of \mathcal{C} are less than or equal to the values of vertices from \mathcal{C} .

$$\max_{u \in N(\mathcal{C})} p(u) \leq \min_{v \in \mathcal{C}} p(v)$$

- Vertex island $\mathcal{C} \subseteq \mathcal{V}$ is a *regular vertex island*, if the stronger condition holds:

$$\max_{u \in N(\mathcal{C})} p(u) < \min_{v \in \mathcal{C}} p(v)$$

Some properties of vertex islands

- The sets of vertices of connected components of vertex-cut at selected level t are regular vertex islands.
- The set $\mathcal{H}_p(\mathcal{N})$ of all regular vertex islands of network \mathcal{N} is a complete hierarchy:
 - two islands are disjoint or one of them is a subset of the other
 - each vertex belongs to at least one island
- Vertex islands are invariant for the strictly increasing transformations of the property p .
- Two linked vertices cannot belong to two disjoint regular vertex islands.

Algorithm for determining maximal regular vertex islands of limited size

- We sink the network into the water, then we lower the water level step by step.
- Each time a new vertex v appears from the water, we check with which of the already visible islands is connected.
- We join these islands and the vertex v obtaining a new (larger) island. These islands are *subislands* of the new island. Vertex v is a *port* of the new island (the vertex with the smallest value).
- This can be done in $\mathcal{O}(\max(n \log n, m))$ time.

algorithm: hierarchy of islands ...

```
islands :=  $\emptyset$   
sort  $\mathcal{V}$  in decreasing order according to  $p$   
for each  $v \in \mathcal{V}$  (in the obtained order) do begin  
  island := new Island()  
  island.port :=  $v$   
  island.subislands :=  $\{i \in \textit{islands} : i \cap N(v) \neq \emptyset\}$   
  islands := islands  $\cup$   $\{\textit{island}\} \setminus \textit{island.subislands}$   
  for each  $i \in \textit{island.subislands}$  do  $i.regular := p(i.port) > p(v)$   
  determine the type of island  
end  
for each  $i \in \textit{islands}$  do  $i.regular := \text{true}$ 
```

... algorithm: list of islands

```
L := ∅  
while islands ≠ ∅ do begin  
  select island ∈ islands  
  islands := islands \ {island}  
  if |island| < min then delete island  
  else if |island| > max ∨ ¬island.regular then begin  
    islands := islands ∪ island.subislands  
    delete island  
  end  
  else L := L ∪ {island}  
end
```

Simple vertex islands

- The set of vertices $\mathcal{C} \subseteq \mathcal{V}$ is a *local vertex peak*, if it is a regular vertex island and all of its vertices have the same value.
- Vertex island with a single local vertex peak is called a *simple vertex island*.
- The types of vertex islands:
 - FLAT – all vertices have the same value
 - SINGLE – island has a single local vertex peak
 - MULTI – island has more than one local vertex peaks
- Only the islands of type FLAT or SINGLE are simple islands.

Determining the type of vertex island

```
if  $|island.subislands| = 0$  then  $island.type := FLAT$   
else if  $|island.subislands| = 1$  then begin  
  select  $i \in island.subislands$   
  if  $i.type \neq FLAT$  then  $island.type := i.type$   
  else if  $p(i.port) = p(v)$  then  $island.type := FLAT$   
  else  $island.type := SINGLE$   
end  
else begin  
  for each  $i \in island.subislands$  do begin  
     $ok := i.type = FLAT \wedge p(i.port) = p(v)$   
    if  $\neg ok$  then break  
  end  
  if  $ok$  then  $island.type := FLAT$   
  else  $island.type := MULTI$   
end
```

Line islands

- The set of vertices $\mathcal{C} \subseteq \mathcal{V}$ is a *line island*, if it is a singleton (degenerated island) or there exists a spanning tree \mathcal{T} on \mathcal{C} such that the values of lines with exactly one endpoint in \mathcal{C} are less than or equal to the values of lines of the tree \mathcal{T} .

$$\max_{\substack{(u;v) \in \mathcal{L}: \\ u \in \mathcal{C} \wedge v \notin \mathcal{C}}} w((u;v)) \leq \min_{e \in \mathcal{L}(\mathcal{T})} w(e)$$

- Line island $\mathcal{C} \subseteq \mathcal{V}$ is a *regular line island*, if the stronger condition holds:

$$\max_{\substack{(u;v) \in \mathcal{L}: \\ u \in \mathcal{C} \wedge v \notin \mathcal{C}}} w((u;v)) < \min_{e \in \mathcal{L}(\mathcal{T})} w(e)$$

Some properties of line islands

- The sets of vertices of connected components of line-cut at selected level t are regular line islands.
- The set $\mathcal{H}_w(\mathcal{N})$ of all nondegenerated regular line islands of network \mathcal{N} is hierarchy (not necessarily complete):
 - two islands are disjoint or one of them is a subset of the other
- Line islands are invariant for the strictly increasing transformations of the weight w .
- Two linked vertices may belong to two disjoint regular line islands.

Algorithm for determining maximal regular line islands of limited size

- We sink the network into the water, then we lower the water level step by step.
- Each time a new line e appears from the water, we check with which of the already visible islands is connected (there are exactly two such islands).
- We join these two islands obtaining a new (larger) island. These islands are *subislands* of the new island. Line e is a *port* of the new island (not necessarily the line with the smallest value).
- This can be done in $\mathcal{O}(m \log n)$ time.

algorithm: hierarchy of islands ...

```

islands :=  $\{\{v\} : v \in \mathcal{V}\}$ 
for each  $i \in \textit{islands}$  do  $i.\textit{port} := \text{null}$ 
sort  $\mathcal{L}$  in decreasing order according to  $w$ 
for each  $e(u;v) \in \mathcal{L}$  (in the obtained order) do begin
     $i1 := \textit{island} \in \textit{islands} : u \in \textit{island}$ 
     $i2 := \textit{island} \in \textit{islands} : v \in \textit{island}$ 
    if  $i1 \neq i2$  then begin
         $\textit{island} := \text{new Island}()$ 
         $\textit{island}.\textit{port} := e$ 
         $\textit{island}.\textit{subisland1} := i1$ 
         $\textit{island}.\textit{subisland2} := i2$ 
         $\textit{islands} := \textit{islands} \cup \{\textit{island}\} \setminus \{i1, i2\}$ 
         $i1.\textit{regular} := i1.\textit{port} = \text{null} \vee w(i1.\textit{port}) > w(e)$ 
         $i2.\textit{regular} := i2.\textit{port} = \text{null} \vee w(i2.\textit{port}) > w(e)$ 
    end
    determine the type of  $\textit{island}$ 
end
for each  $i \in \textit{islands}$  do  $i.\textit{regular} := \text{true}$ 

```

... algorithm: list of islands

```
L := ∅  
while islands ≠ ∅ do begin  
  select island ∈ subislands  
  subislands := subislands \ {island}  
  if |island| < min then delete island  
  else if |island| > max ∨ ¬island.regular then begin  
    islands := islands ∪ {island.subisland1, island.subisland2}  
    delete island  
  end  
  else L := L ∪ {island}  
end
```

Simple line islands

- The set of vertices $\mathcal{C} \subseteq \mathcal{V}$ is a *local line peak*, if it is a regular line island and there exists a spanning tree of the corresponding induced network, in which all lines have the same value as the line with the largest value.
- Line island with a single local line peak is called a *simple line island*.
- The types of line islands:
 - FLAT – there exists a spanning tree, in which all lines have the same value as the line with the largest value.
 - SINGLE – island has a single local line peak.
 - MULTI – island has more than one local line peaks.
- Only the islands of type FLAT or SINGLE are simple islands.

Determining the type of line islands

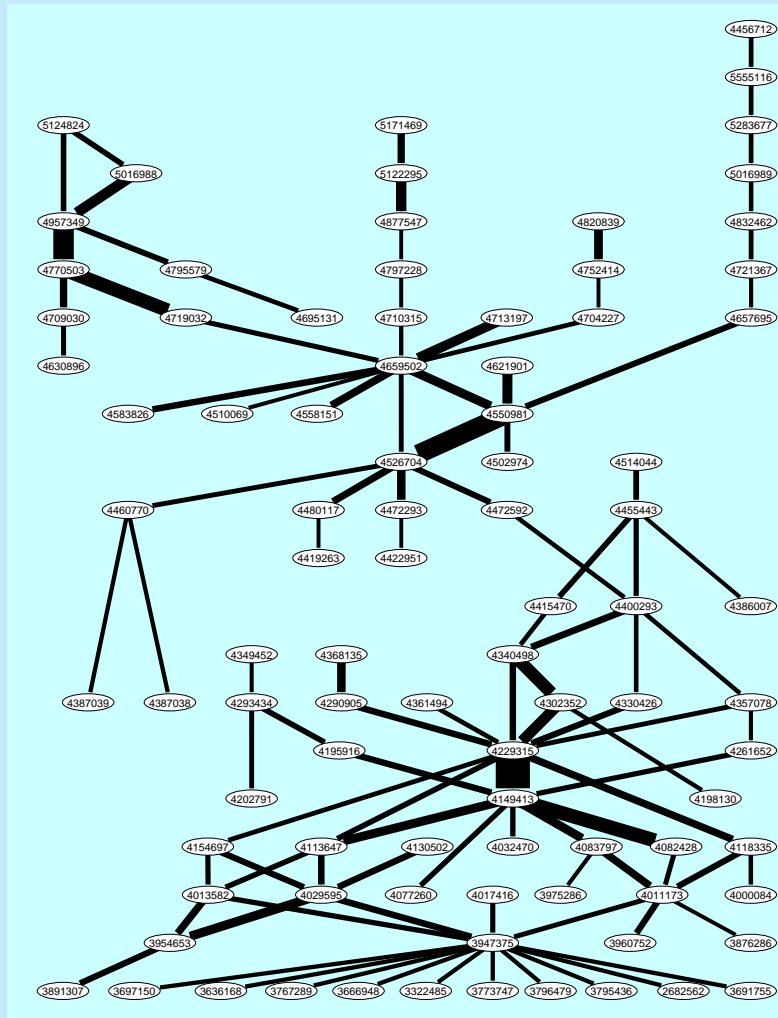
$$p1 := i1.type = \text{FLAT} \wedge (i1.port = \mathbf{null} \vee w(i1.port) = w(e))$$
$$p2 := i2.type = \text{FLAT} \wedge (i2.port = \mathbf{null} \vee w(i2.port) = w(e))$$

if $p1 \wedge p2$ **then** $island.type := \text{FLAT}$

else if $p1 \vee p2$ **then** $island.type := \text{SINGLE}$

else $island.type := \text{MULTI}$

Example: US Patents



The citation network of US patents from 1963 to 1999 (<http://www.nber.org/patents/>) is an example of very large network (3774768 vertices and 16522438 arcs) that, using some special options in **Pajek**, can still be analyzed on PC with at least 1 G memory. The islands algorithm was applied on Hummon-Doreian SPC weights.

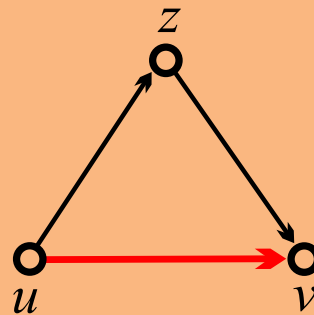
The obtained main island is presented in the figure. The vertices represent patents, the size of a line is proportional to its weight. Collecting from the *United States Patent and Trademark Office* (<http://patft.uspto.gov/netahtml/srchnum.htm>) the basic data about the patents we can see that they deal with the 'liquid crystal displays'.

Example: The Edinburgh Associative Thesaurus

- The Edinburgh Associative Thesaurus is a set of words and the counts of word associations as collected from subjects.
- The data were collected by asking several people to say a word which first comes to their mind upon receiving the stimulus word.
- The network contains 23219 vertices (words) and 325624 arcs (stimulus→response), including 564 loops. Almost 70% of arcs have value 1.
- The subjects were mostly undergraduates from a wide variety of British universities. The age range of the subjects was from 17 to 22 with a mode of 19. The sex distribution was 64 per cent male and 36 per cent female. The data were collected between June 1968 and May 1971.

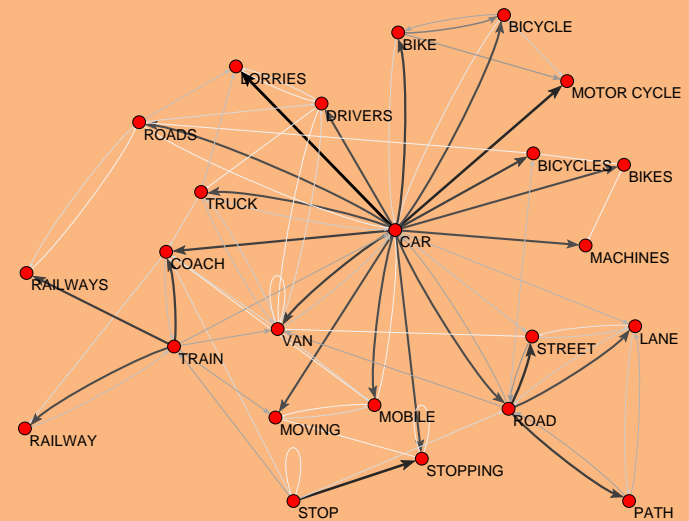
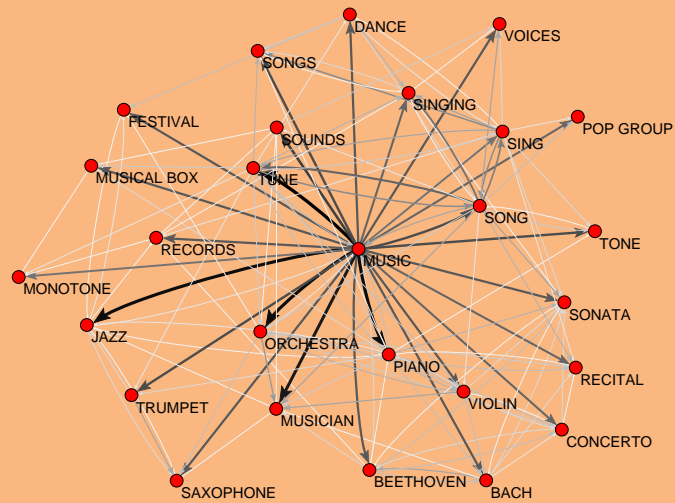
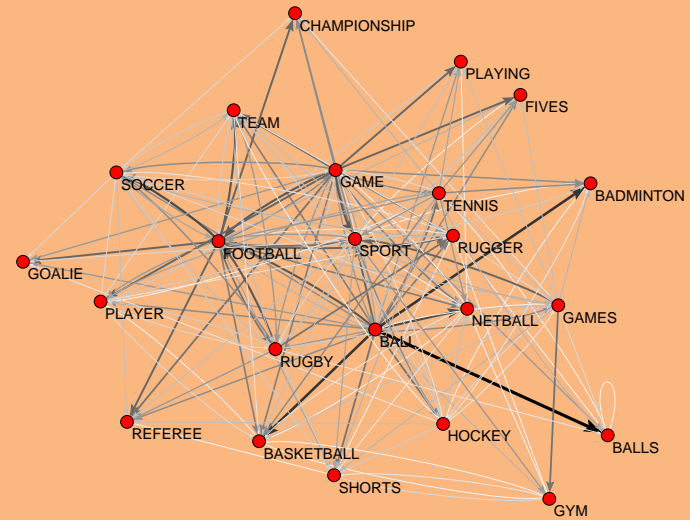
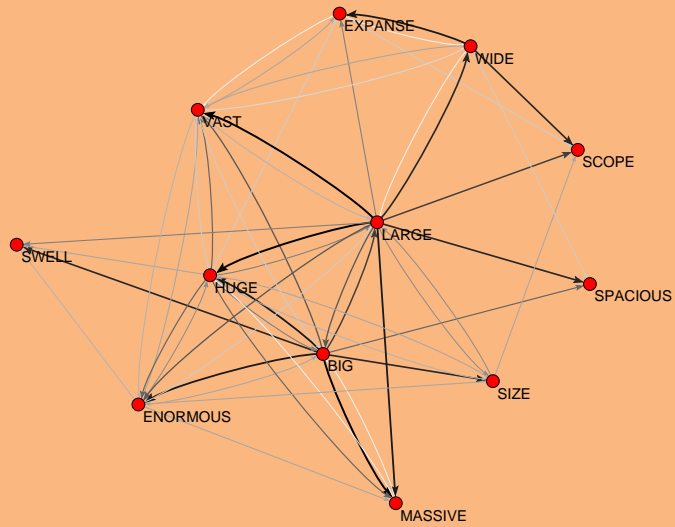
Transitivity weight

- We would like to identify the most important themes – groups of words with the strongest ties.
- For each arc we determined its weight by counting, to how many transitive triangles it belongs (we are also interested in indirect ties).

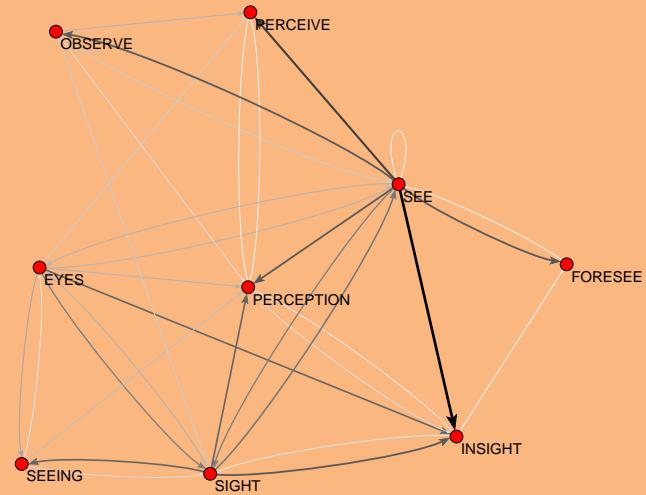
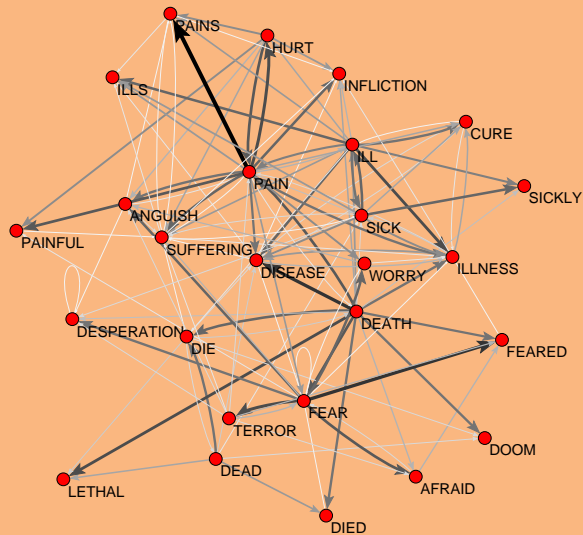
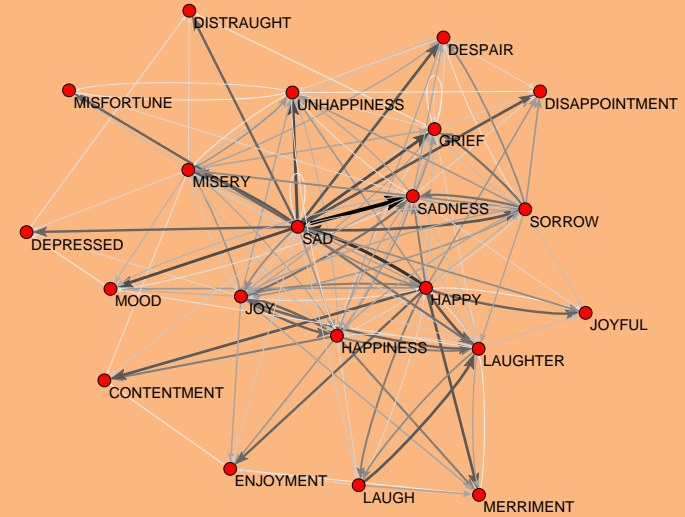
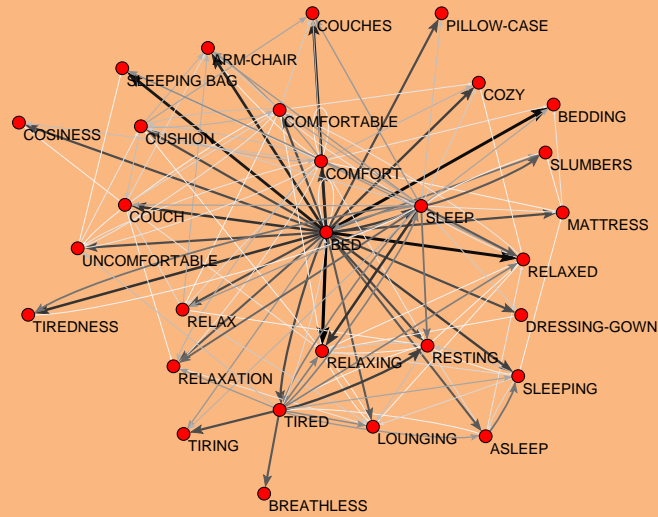


- There are 53 line islands of size at least 5 and at most 30. They contain 664 vertices (all together).

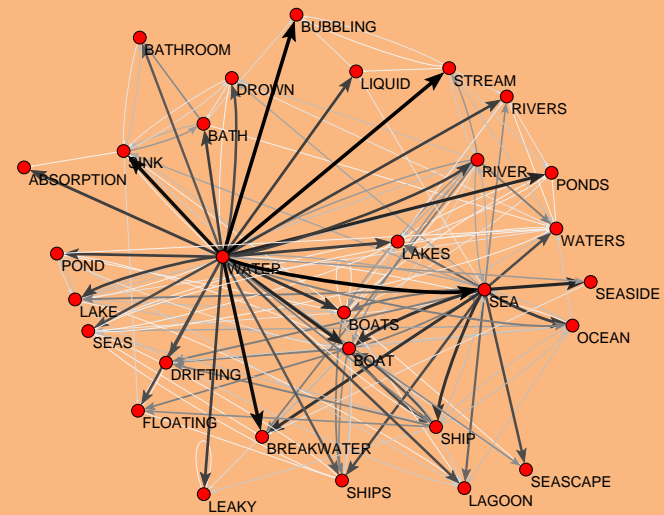
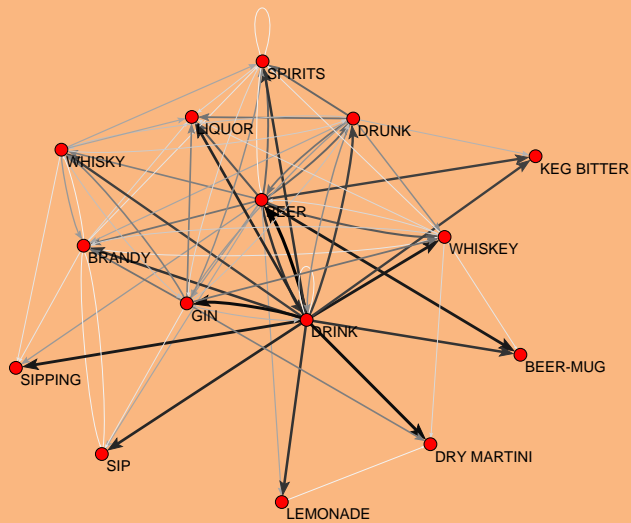
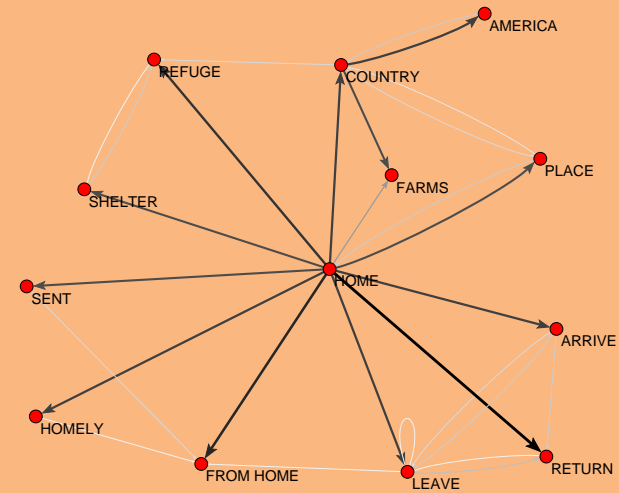
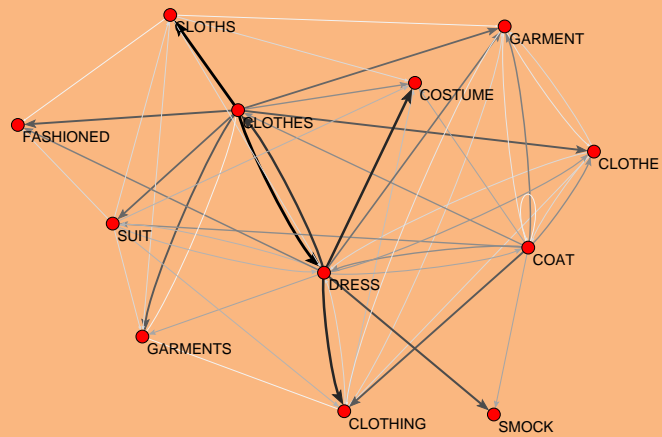
Selected themes in EAT



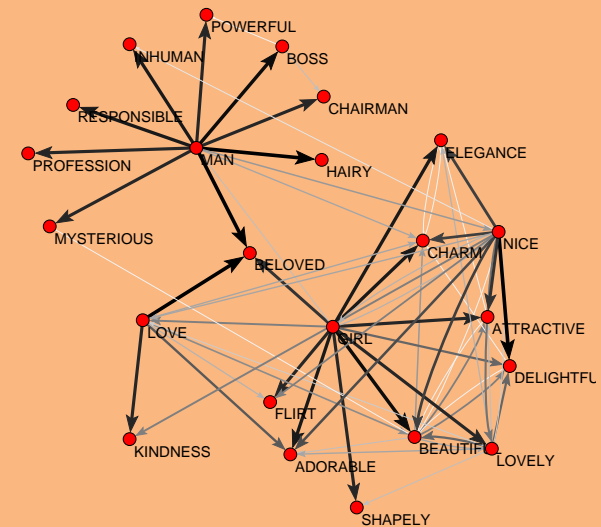
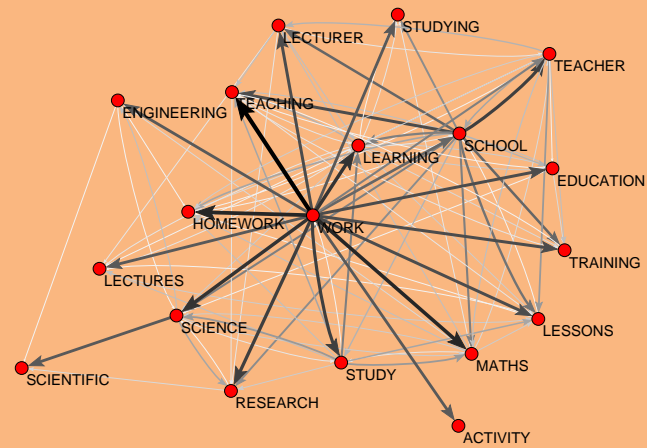
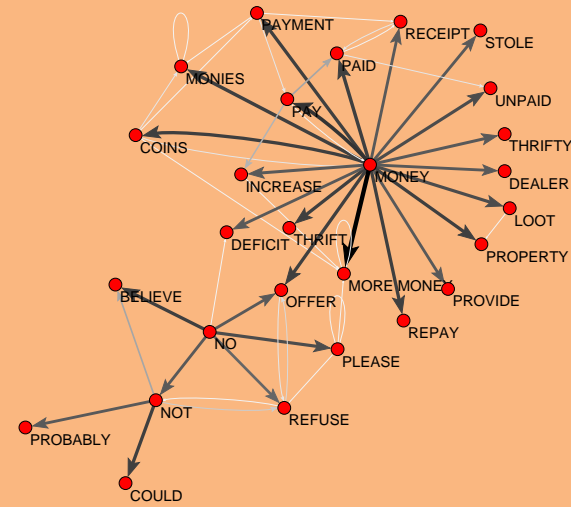
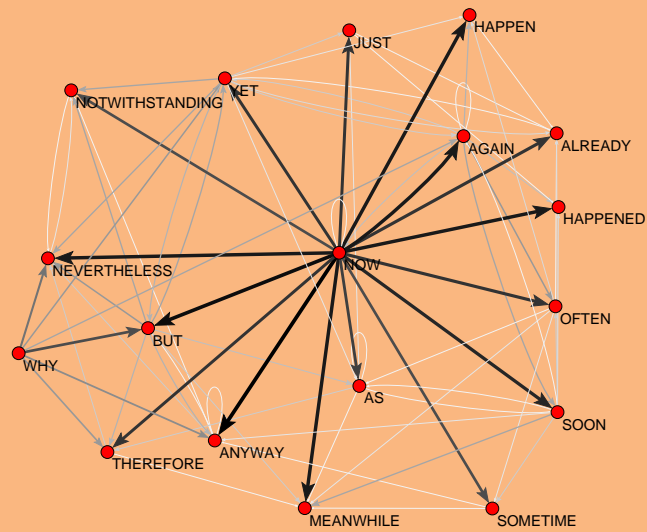
Selected themes in EAT



Selected themes in EAT



Selected themes in EAT



Example: Amazon CDs and books networks

The *vertices* in Amazon networks are books / CDs; while the *ares* are determined based on the list of products (CDs/books) under the title: Customers who bought this CD/book also bought

SEARCH
Books

GO!

WEB SEARCH
Powered by A9.com

BOOK INFORMATION
[buying info](#)
[editorial reviews](#)
[customer reviews](#)
[search inside](#)
[reading group guide](#)

RECENTLY VIEWED
[The Name of the Rose: Including Postscript to the Name of the Rose](#) by Umberto Eco
[All My Hits - Todos Mis Exitos](#) ~ Selena
[Dreaming of You](#) ~ Selena
[Selena - Greatest Hits](#) ~ Selena

See more in the Page You Made

Featured Item:

[Selena](#) ~ Various Artists

RATE THIS BOOK
[\(Sign in to rate this item.\)](#)

The Da Vinci Code (Random House Large Print) [LARGE PRINT]
 by [DAN BROWN](#)

READY TO BUY?
 Quantity: 1
[Add to Shopping Cart](#)
 or
[Sign in](#) to turn on 1-Click ordering.

MORE BUYING CHOICES
[57 used & new](#) from \$9.95
 Have one to sell? [Sell yours here](#)

[Add to Wish List](#)
 Don't have one? We'll set one up for you.
[Add to Wedding Registry](#)

SEARCH INSIDE!

 Search inside another edition of this book

List Price: \$26.95
Price: \$26.95 & This item ships for **FREE with Super Saver Shipping.** See details.
Availability: Usually ships within 24 hours

57 used & new from \$9.95

Edition: Hardcover
Other Editions: List Price: Our Price: Other Offers:
 Hardcover (1st ed) ~~\$24.95~~ \$14.97 **305 used & new** from \$5.49
 Audio Cassette (Abridged) ~~\$29.95~~ \$18.16 **57 used & new** from \$12.00
 Audio Cassette (Unabridged) ~~\$29.95~~ \$27.17 **48 used & new** from \$22.95
[See all 9 editions...](#)

[See more product details](#)

Better Together
 Buy this book with [Angels & Demons](#) by Dan Brown today!
Buy Together Today: \$32.54
[Buy both now!](#)

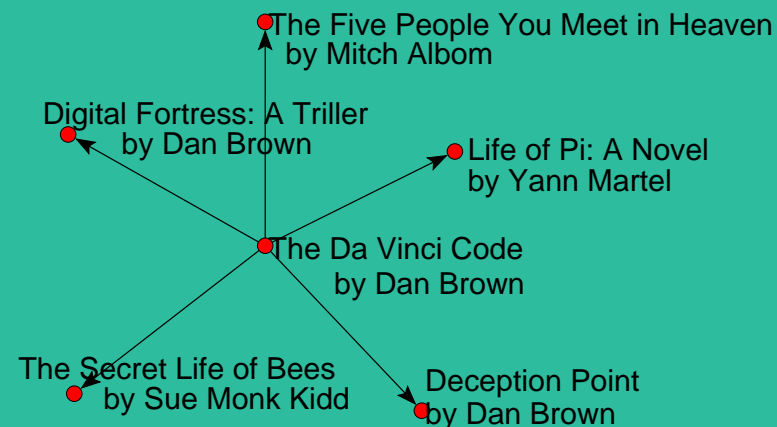
Customers who bought this book also bought:

- [The Five People You Meet in Heaven](#) by Mitch Albom [\(Rate it\)](#)
- [Life of Pi](#) by Yann Martel [\(Rate it\)](#)
- [Deception Point](#) by Dan Brown [\(Rate it\)](#)
- [The Secret Life of Bees](#) by Sue Monk Kidd [\(Rate it\)](#)
- [Digital Fortress: A Thriller](#) by Dan Brown [\(Rate it\)](#)

Explore Similar Items: [19 in Books](#), [20 in Music](#), and [20 in DVD](#)

Product Details

- Hardcover:** 752 pages ; Dimensions (in inches): 1.51 x 9.58 x 6.42



... Amazon CDs and books networks

The screenshot shows the Amazon product page for Selena's album "All My Hits - Todos Mis Exitos". The page includes a search bar, a "READY TO BUY?" section with pricing and availability information, a "MUSIC INFORMATION" section with links to explore the album, a "RECENTLY VIEWED" section with recommendations, a "Customers who bought this title also bought:" section, a "Customers who viewed this album also viewed:" section, and a "Product Details" section with technical specifications and a customer review.

SEARCH
Popular Music

WEB SEARCH

Powered by A9.com

All My Hits - Todos Mis Exitos
Selena

READY TO BUY?
13 used & new from \$4.99

Have one to sell?

Don't have one?
We'll set one up for you.

Availability: Usually ships in 1-2 business days
13 used & new from \$4.99

[see larger picture](#)

MUSIC INFORMATION
Explore this album
buying info
[listen to samples](#)
[editorial reviews](#)
[customer reviews](#)

RECENTLY VIEWED
 ~ Selena [\(Rate it\)](#)
 ~ Selena [\(Rate it\)](#)
 ~ Jennifer Lopez [\(Rate it\)](#)
 ~ Jimmy Smith [\(Rate it\)](#)

Customers who bought this title also bought:

- [All My Hits - Todos Mis Exitos, Vol. 2 \[LIMITED EDITION\]](#) ~ Selena
- [Mis Mejores Canciones: 17 Super Exitos](#) ~ Selena [\(Rate it\)](#)

Customers who viewed this album also viewed:

- [Selena - Greatest Hits](#) ~ Selena [\(Rate it\)](#)
- [Exitos Y Recuerdos](#) ~ Selena [\(Rate it\)](#)
- [Anthology](#) ~ Selena [\(Rate it\)](#)
- [Ones \(CD & DVD\) \(Limited Edition Package\)](#) ~ Selena [\(Rate it\)](#)
- [Selena - Greatest Hits \(Bonus DVD\)](#) ~ Selena [\(Rate it\)](#)
- [Momentos Intimos](#) ~ Selena [\(Rate it\)](#)
- [Live, The Last Concert - - February 26, 1995, Houston TX](#) ~ Selena [\(Rate it\)](#)

Explore Similar Items: 9 in Music, 5 in DVD, and 1 in Books

Product Details

- Essential recordings:** [Selena](#)
- Audio CD** (March 9, 1999)
- Original Release Date:** March 9, 1999
- Number of Discs:** 1
- Label:** Emi International
- Catalog:** #97886
- ASIN:** B0000017XB
- Average Customer Review:** ★★★★★ Based on 48 reviews. [Write a review.](#)
- Amazon.com Sales Rank:** 8,528

Featured Item:

[Selena](#) ~ Various Artists

RATE THIS ALBUM
[\(Sign in to rate this item.\)](#)

Using relatively simple program written in Python we 'harvested' the books network from June 16 till June 27, 2004; and the CDs network from July 7 till July 23, 2004.

We harvested only the portion of each network reachable from the selected starting book/CD. The books network has 216737 vertices and 982296 arcs.

The CDs network has 79244 vertices and 526271 arcs.

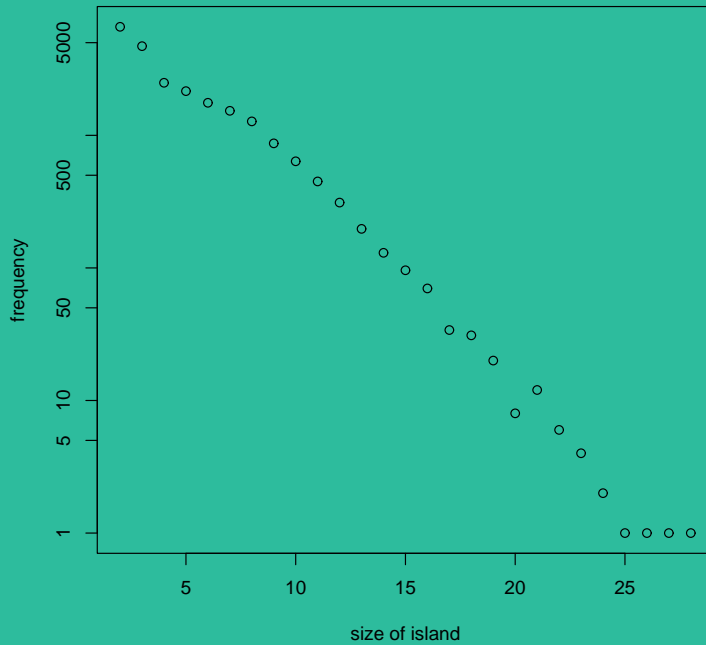
By the construction both networks have limited out-degree and are weakly connected. 178281 books have the out-degree 5; and 55373 CDs have out-degree 8.

The networks were analysed by *Nataša Kejžar* and *Simona Korenjak-Černe*.

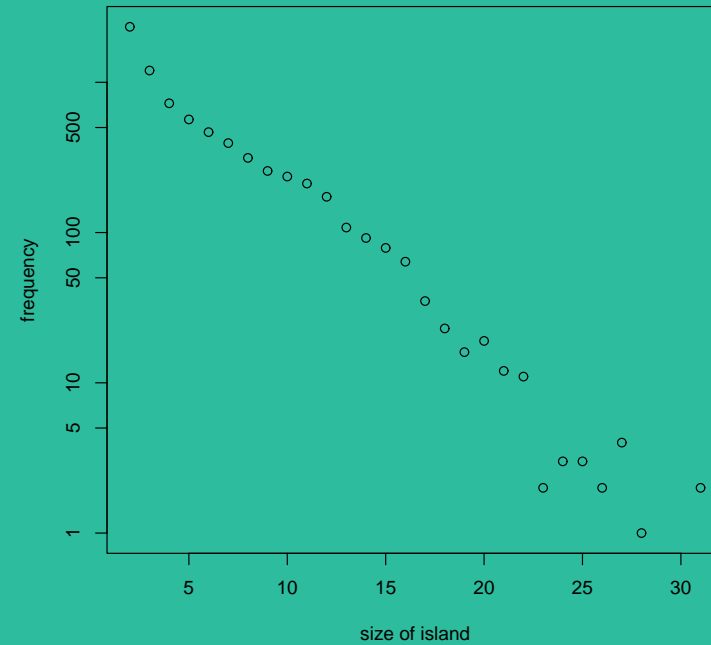
Simple arc islands size distribution

We took the number of cyclic triangles as weights on arcs.

Books' network – islands distribution

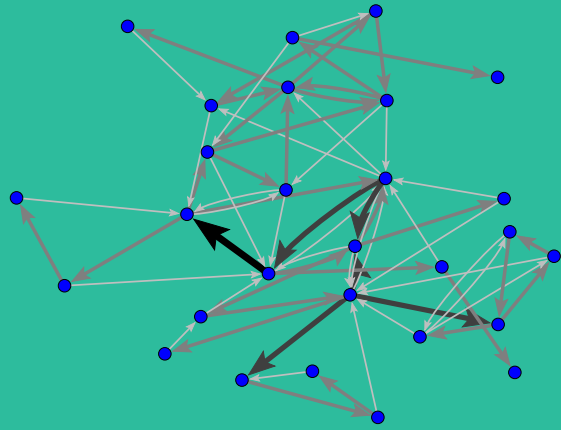


CDs' network – islands distribution

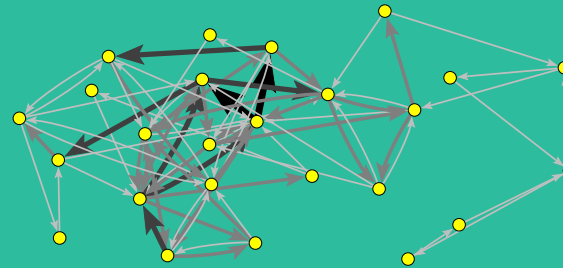


Islands with at least 25 vertices

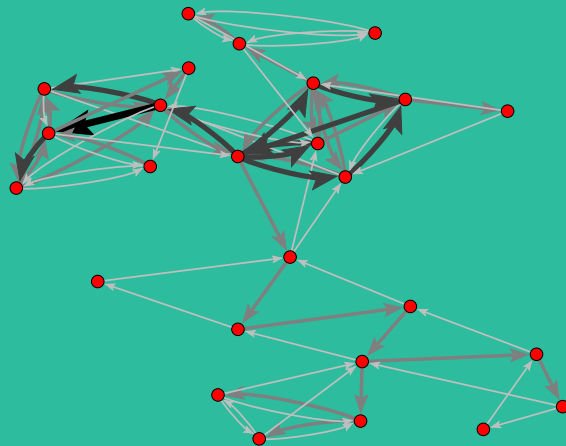
.NET programming, programming in C#



Catherine Cookson novels

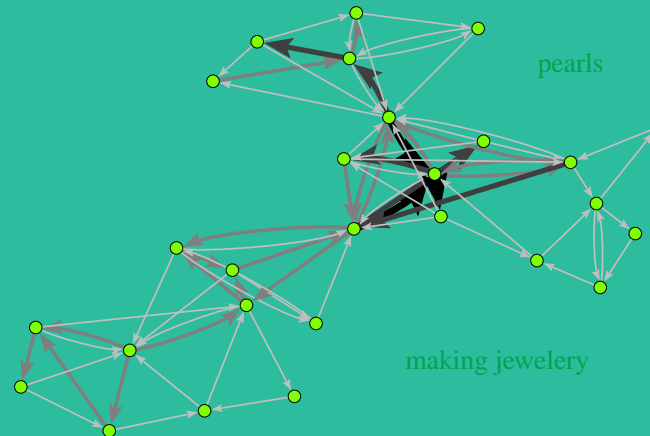


near death experience



after death, across the unknown

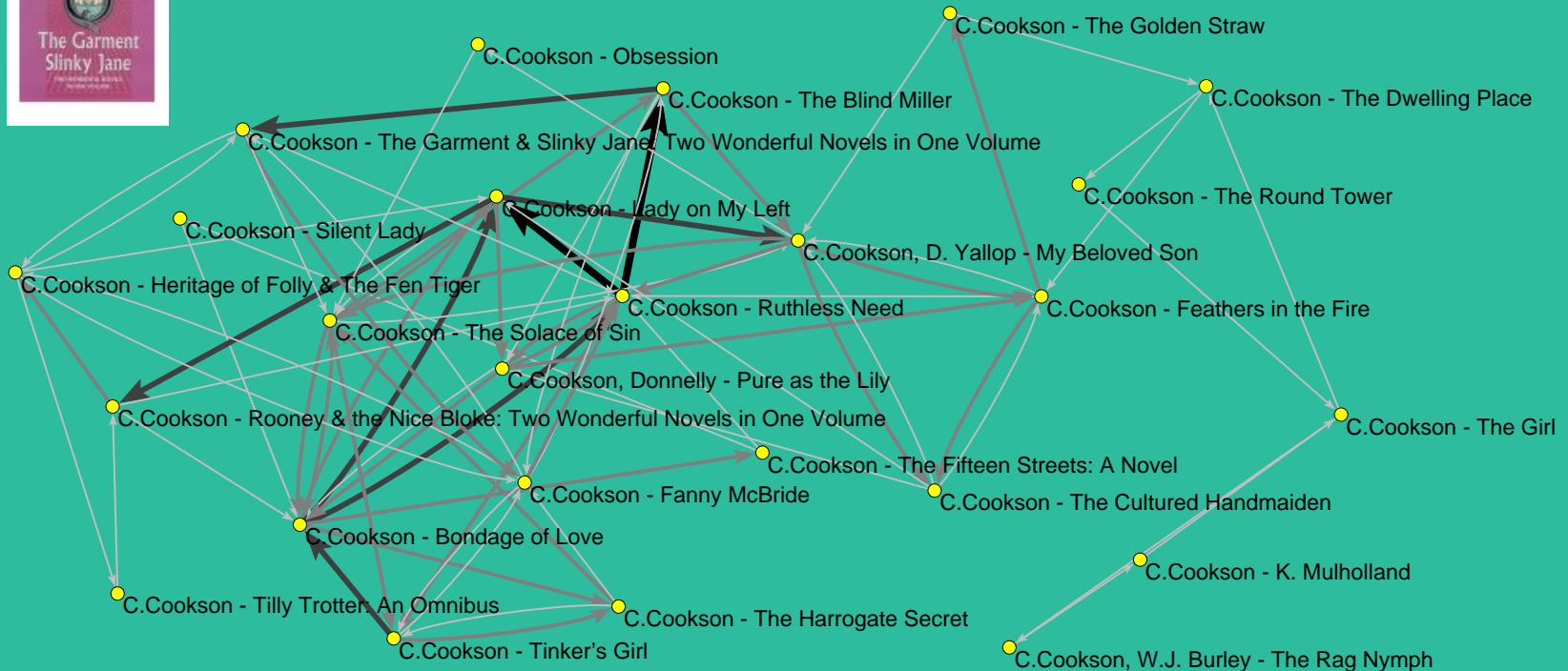
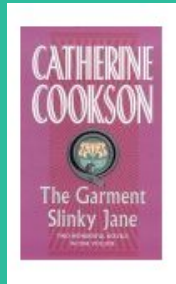
pearls



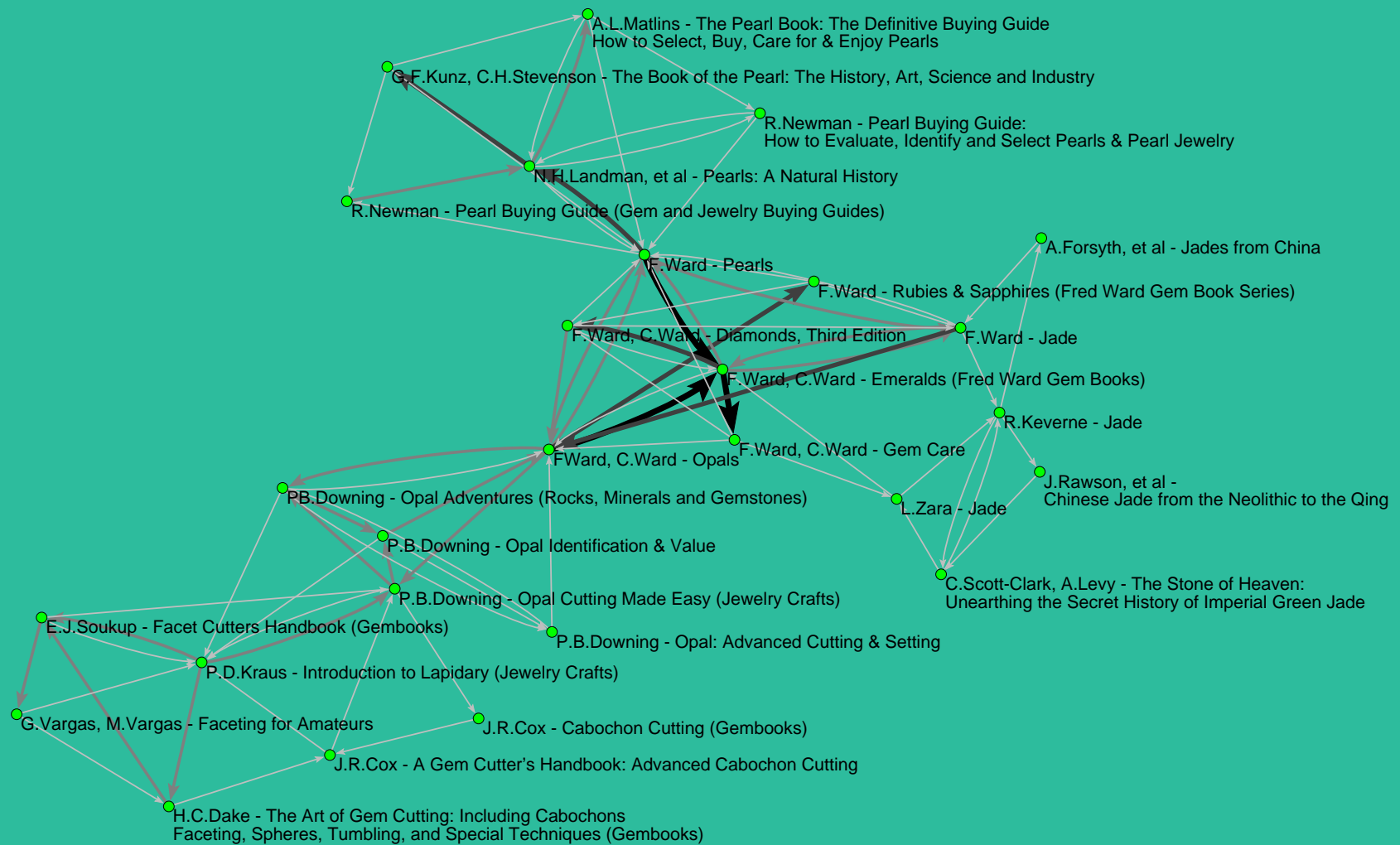
all gems

making jewellery

Island of Catherine Cookson novels



Island of precious stones



Conclusions

- We proposed an approach to the analysis of networks that can be used also for very large networks with millions of vertices and lines.
- The proposed approach is very general – it can be applied to any property of vertices (vertex islands) and to any weight on lines (line islands).
- The islands algorithms are implemented in **Pajek** – a program (for Windows) for large network analysis and visualization

<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>

They are available also as a separate program at

<http://vlado.fmf.uni-lj.si/pub/networks/>

- The last version of these slides is available at

<http://vlado.fmf.uni-lj.si/pub/networks/doc/mix/islands.pdf>