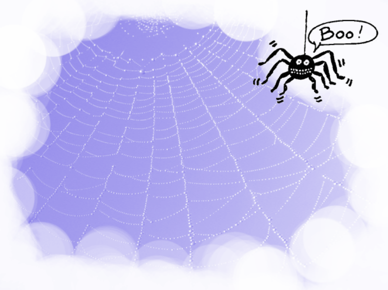
Mining Social and Semantic Network Data on the Web



Markus Schatten, PhD

University of Zagreb
Faculty of Organization and Informatics

May 4, 2011

Introduction

Web 2.0, Semantic Web, Web 3.0

Network science

Data collection

- Semistructured data
- XPath
- Regular expressions
- Ajax
- APIs

Data analysis

- Network matrices
- Folksonomy model
- Logic programming

Case study - CROSBI

Outline

Questions

- Why mining social and semantic network data?
- How to collect these data?
- What are the common obstacles and how to solve them?
- How to analyze these data?

Outline

Web 2.0, Semantic Web, Web 3.0

+

=

Why should one bother with these buzzwords?

Why mining (social) web data?

Millions of people,

Why mining (social) web data?

Millions of people, on thousands of sites,

Why mining (social) web data?

Millions of people, on thousands of sites, across hundreds of diverse cultures,

Why mining (social) web data?

Millions of people, on thousands of sites, across hundreds of diverse cultures, leave “trails”

Why mining (social) web data?

Millions of people, on thousands of sites, across hundreds of diverse cultures, leave “trails” about various aspects of their lives...

Why mining (social) web data?

Millions of people, on thousands of sites, across hundreds of diverse cultures, leave "trails" about various aspects of their lives...

"Social systems are systems of communication"

(Niklas Luhman)

Why mining (social) web data?

Millions of people, on thousands of sites, across hundreds of diverse cultures, leave “trails” about various aspects of their lives...

“Social systems are systems of communication”

(Niklas Luhman)

By analyzing the “trails” (the results of social systems structural coupling) we can analyze the social system itself:

Why mining (social) web data?

Millions of people, on thousands of sites, across hundreds of diverse cultures, leave “trails” about various aspects of their lives...

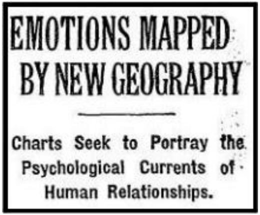
“Social systems are systems of communication”

(Niklas Luhman)

By analyzing the “trails” (the results of social systems structural coupling) we can analyze the social system itself:politics, culture, media, business, technology, science ...

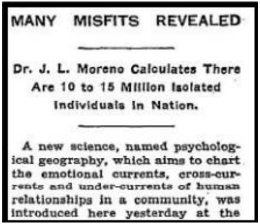
Outline

The “new” science of networks



EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

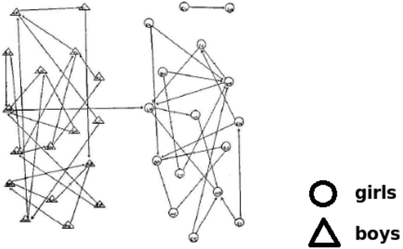


MANY MISFITS REVEALED

Dr. J. L. Moreno Calculates There Are 10 to 15 Million Isolated Individuals in Nation.

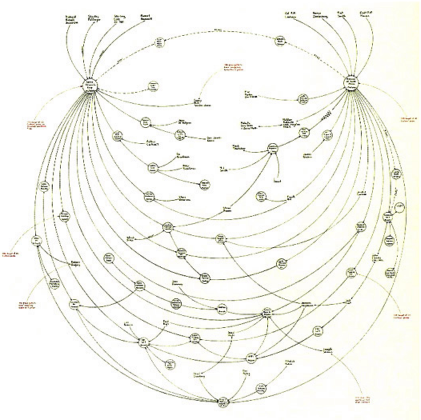
A new science, named psychological geography, which aims to chart the emotional currents, cross-currents and under-currents of human relationships in a community, was introduced here yesterday at the

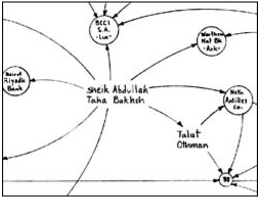
source: The New York Times (April 3rd, 1933., p. 17).



source: An Attraction Network in a Fourth Grade Class (Moreno, Who shall survive?, 1934).

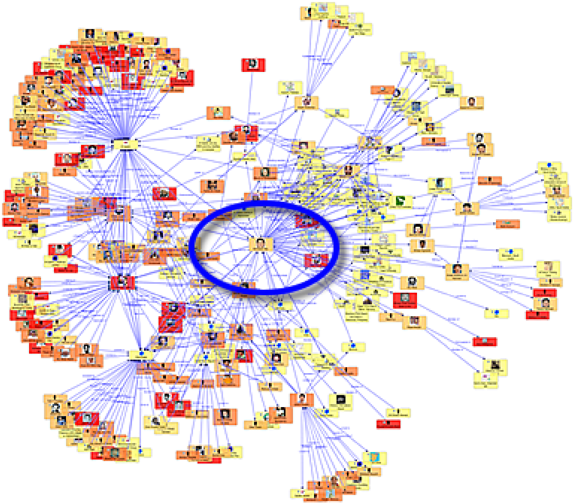
Political and financial networks



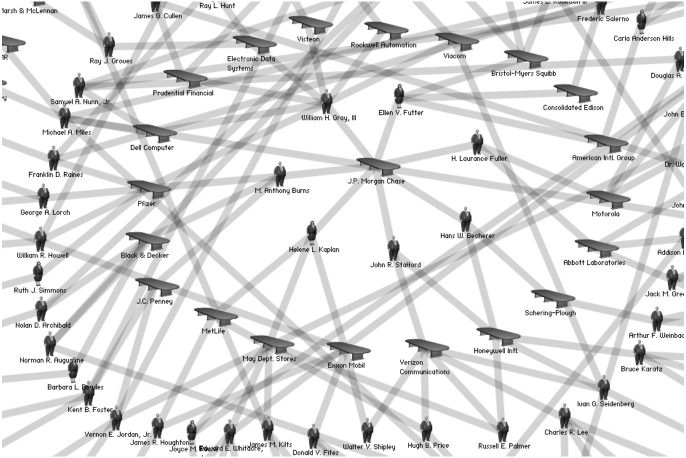


Mark Lombardi (1980s & 1990s)

Terrorist networks

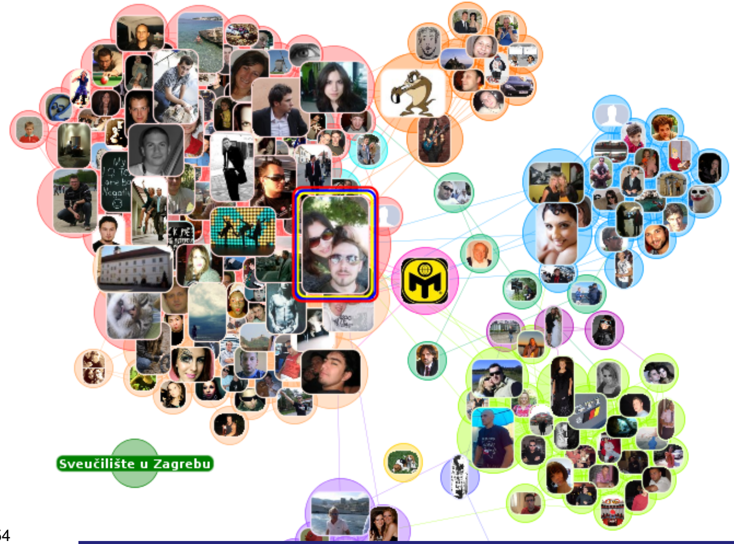


Board of directors membership networks

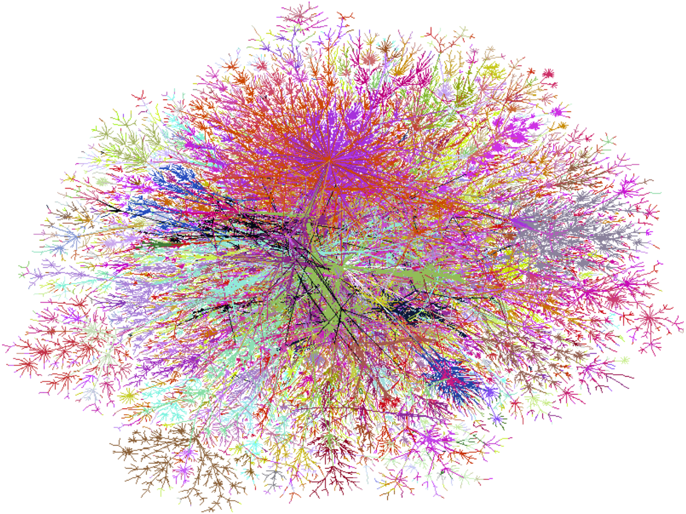


source: `http://theyrule.net`

On-line social networks

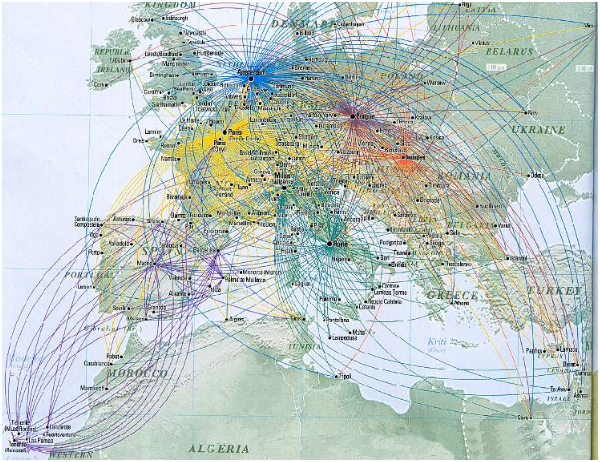


Internet



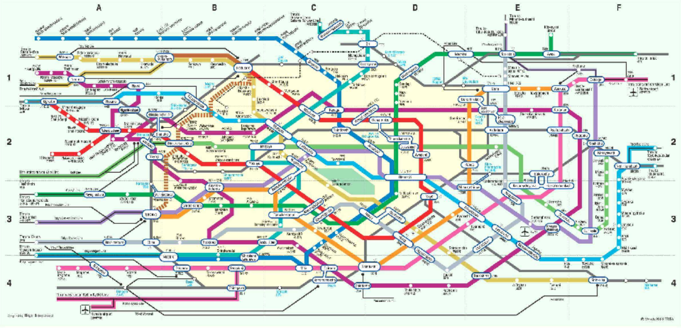
source: Bill Cheswick `http://www.cheswick.com/ches/map/gallery/index.html`

Air traffic networks



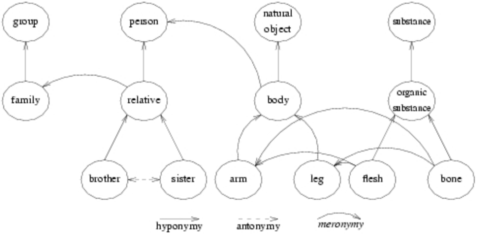
source: Northwest Airlines WorldTraveler Magazine

Railway networks



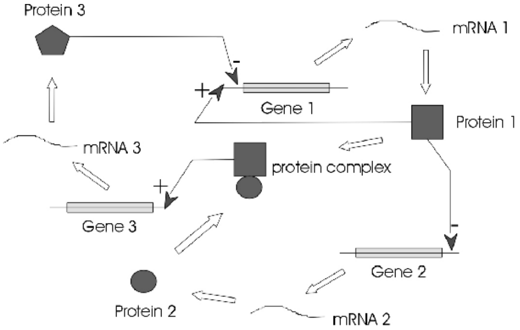
source: TRTA, March 2003 - Tokyo rail map

Semantic networks



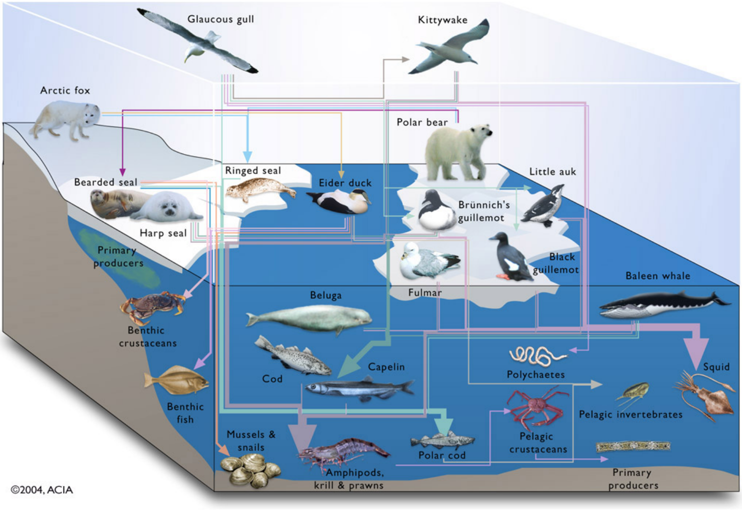
source: `http://wordnet.princeton.edu/man/wnlicens.7WN`

Gene networks



source: `http://www.zaik.uni-koeln.de/bioinformatik/regulatorynets.html.en`

Food chain networks



On-line social and semantic networks

Where can we find networks?

- social networking sites

On-line social and semantic networks

Where can we find networks?

- social networking sites
- forums, buletinboards, discussion groups, mailing lists

On-line social and semantic networks

Where can we find networks?

- social networking sites
- forums, buletinboards, discussion groups, mailing lists
- blogs, microblogs, vlogs

On-line social and semantic networks

Where can we find networks?

- social networking sites
- forums, buletinboards, discussion groups, mailing lists
- blogs, microblogs, vlogs
- wikis, semantic wikis

On-line social and semantic networks

Where can we find networks?

- social networking sites
- forums, buletinboards, discussion groups, mailing lists
- blogs, microblogs, vlogs
- wikis, semantic wikis
- social tagging

On-line social and semantic networks

Where can we find networks?

- social networking sites
- forums, buletinboards, discussion groups, mailing lists
- blogs, microblogs, vlogs
- wikis, semantic wikis
- social tagging
- RSS feeds, RDF repositories, metadata collections

On-line social and semantic networks

Where can we find networks?

- social networking sites
- forums, buletinboards, discussion groups, mailing lists
- blogs, microblogs, vlogs
- wikis, semantic wikis
- social tagging
- RSS feeds, RDF repositories, metadata collections...

On-line social and semantic networks

Where can we find networks?

- social networking sites
- forums, buletinboards, discussion groups, mailing lists
- blogs, microblogs, vlogs
- wikis, semantic wikis
- social tagging
- RSS feeds, RDF repositories, metadata collections...

A social network can be found on any application where users communicate.

On-line social and semantic networks

Where can we find networks?

- social networking sites
- forums, buletinboards, discussion groups, mailing lists
- blogs, microblogs, vlogs
- wikis, semantic wikis
- social tagging
- RSS feeds, RDF repositories, metadata collections...

A social network can be found on any application where users communicate.

A semantic network can be found in any specification where concepts are meaningfully connected.

Outline

Introduction

Web 2.0, Semantic Web, Web 3.0

Network science

Data collection

- Semistructured data
- XPath
- Regular expressions
- Ajax
- APIs

Data analysis

Case study - CROSBI

Common obstacles in data collection

- Semi-structured data (HTML, XHTML, XML, RDF, OWL ...)

```html
<div id="menu">
        <div id="menuitem">

                <form action="login.php" method="post">
                <font color="red"></font><br />
                <strong>User Name:</strong><br />
                <input maxlength="30" name="login_user_name" size="10" tabindex="1" type="text" /><br /><br />
                <strong>Password:</strong><br />
                <input maxlength="30" name="login_password" size="10" tabindex="2" type="password" /><br /><br />
                <input name="submit" tabindex="4" type="submit" value="Login" />
                </form>
                <a href="register.php" title="Register">New user</a><br />
                <a href="lost_pass.php" title="Lost password!">Lost password!</a></div>
                <div id="menuitem"><b>Browse</b><br />
                <div id="menuitem">
                        <a href="project.php?browse=pro">Projects</a>
                </div>
                <div id="menuitem">
                        <a href="project.php?browse=org">Organizations</a>
                </div>
                <div id="menuitem">
                        <a href="userlist.php">Users</a>
                </div>
                <div id="menuitem">
                        <a href="search.php">Search</a>
                </div>
        </div>
```

Common obstacles in data collection

- Semi-structured data (HTML, XHTML, XML, RDF, OWL ...)

```html
<div id="menu">
        <div id="menuitem">

                <form action="login.php" method="post">
                <font color="red"></font><br />
                <strong>User Name:</strong><br />
                <input maxlength="30" name="login_user_name" size="10" tabindex="1" type="text" /><br /><br />
                <strong>Password:</strong><br />
                <input maxlength="30" name="login_password" size="10" tabindex="2" type="password" /><br /><br />
                <input name="submit" tabindex="4" type="submit" value="Login" />
                </form>
                <a href="register.php" title="Register">New user</a><br />
                <a href="lost_pass.php" title="Lost password!">Lost password!</a></div>
                <div id="menuitem"><b>Browse</b><br />
                <div id="menuitem">
                        <a href="project.php?browse=pro">Projects</a>
                </div>
                <div id="menuitem">
                        <a href="project.php?browse=org">Organizations</a>
                </div>
                <div id="menuitem">
                        <a href="userlist.php">Users</a>
                </div>
                <div id="menuitem">
                        <a href="search.php">Search</a>
                </div>
        </div>
```

- The need for pattern matching

TOP Administration (2009-11-26 13:02:10) - **Markus Schatten**
ompare to previous (23339)

TOP Administration (2011-01-30 05:47:00) - **Markus Schatten**
ompare to previous (25144)

TOP Administration (2011-01-30 05:48:26) - **Markus Schatten**
ompare to previous (44477)

Common obstacles in data collection

- Semi-structured data (HTML, XHTML, XML, RDF, OWL ...)

```html
<div id="menu">
        <div id="menuitem">

                <form action="login.php" method="post">
                <font color="red"></font><br />
                <strong>User Name:</strong><br />
                <input maxlength="30" name="login_user_name" size="10" tabindex="1" type="text" /><br /><br />
                <strong>Password:</strong><br />
                <input maxlength="30" name="login_password" size="10" tabindex="2" type="password" /><br /><br />
                <input name="submit" tabindex="4" type="submit" value="Login" />
                </form>
                <a href="register.php" title="Register">New user</a><br />
                <a href="lost_pass.php" title="Lost password!">Lost password!</a></div>
                <div id="menuitem"><b>Browse</b><br />
                <div id="menuitem">
                        <a href="project.php?browse=pro">Projects</a>
                </div>
                <div id="menuitem">
                        <a href="project.php?browse=org">Organizations</a>
                </div>
                <div id="menuitem">
                        <a href="userlist.php">Users</a>
                </div>
                <div id="menuitem">
                        <a href="search.php">Search</a>
                </div>
        </div>
```

- The need for pattern matching

TOP Administration (2009-11-26 13:02:10) - **Markus Schatten**
ompare to previous (23339)

TOP Administration (2011-01-30 05:47:00) - **Markus Schatten**
ompare to previous (25144)

TOP Administration (2011-01-30 05:48:26) - **Markus Schatten**
ompare to previous (44477)

- Client-side content generation

Aston V. - Stoke City

(archive pre-match odds)

Saturday, 23 Apr 2011, 16:00

Final result 1:1 (1:1,0:0)

→ Show all "Premier League" matches

Hidden bookmakers (register or log in to add any of them to comparison):

after a few seconds

Aston V. - Stoke City

(archive pre-match odds)

Saturday, 23 Apr 2011, 16:00

Final result 1:1 (1:1,0:0)

1X2 | AH | O/U | DNB | DC | CS | HT/FT | O/E

Full Time | 1st Half | 2nd Half

Bookmakers	1	X	2	Payout
10Bet	2.10	3.19	3.78	94.9%
12BET	2.20	3.06	3.72	95.2%
188bet	2.21	3.15	3.85	97.1%
5Dimes	2.21	3.23	3.92	98.3%
AllYouBet	2.10	3.19	3.78	94.9%
bet-at-home	2.05	3.10	3.70	92.5%

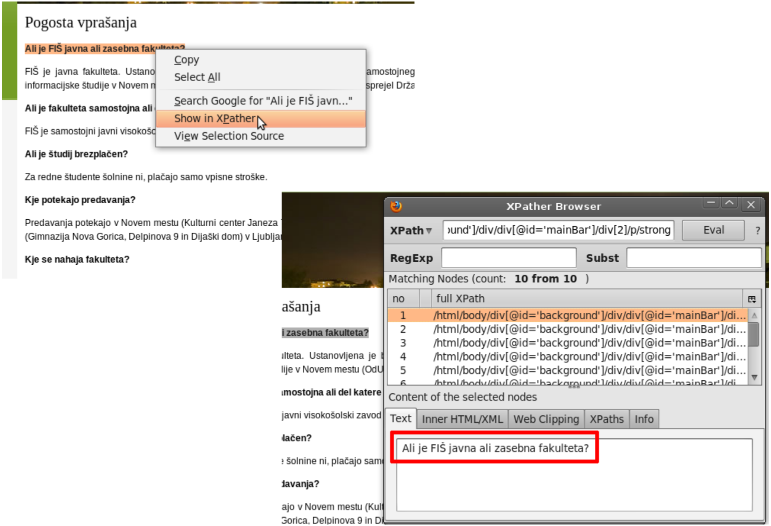
Semistructured data and XPath

- query language for semistructured data
- part of the W3C standard
- uses path expressions to navigate through documents
- has a standard library with over 100 useful functions

Example:

```
/html/body/div[@id='background']/div/div[@id='mainBar']/div/p[last()-1]/
```

XPather (Firefox add-on)



Regular expressions

- Useful for extracting various formatted data and searching for keywords

Example (date matching)

```
27-9-2001    [0-9]{1,2}[0-9]{1,2}[1-2][0-9]{3}
```

Client side generated content

- Some sites use Ajax to generate data on page load or user interaction (e.g. mouse click, mouse over etc.)

Client side generated content

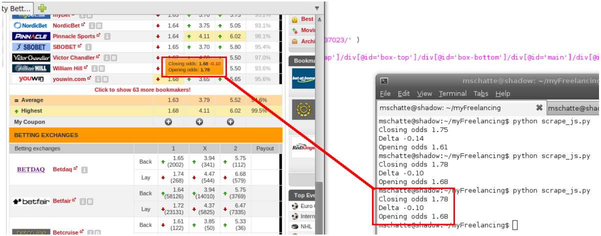
- Some sites use Ajax to generate data on page load or user interaction (e.g. mouse click, mouse over etc.)
- In combination with obfuscated JavaScript code this is hard to scrape!
- Current solutions:
  - Browser scripting (can be slow and cumbersome) - e.g. with selenium

Client side generated content

- Some sites use Ajax to generate data on page load or user interaction (e.g. mouse click, mouse over etc.)
- In combination with obfuscated JavaScript code this is hard to scrape!
- Current solutions:
  - Browser scripting (can be slow and cumbersome) - e.g. with selenium
  - Headless browser - e.g. mechanize, spidermonkey, HtmlUnit

Client side generated content

- Some sites use Ajax to generate data on page load or user interaction (e.g. mouse click, mouse over etc.)
- In combination with obfuscated JavaScript code this is hard to scrape!
- Current solutions:
  - Browser scripting (can be slow and cumbersome) - e.g. with selenium
  - Headless browser - e.g. mechanize, spidermonkey, HtmlUnit



Open APIs

- Most popular social web sites have an open API that allow you to connect your application to their site (Facebook, Twitter, Wikipedia, ...)

Open APIs

- Most popular social web sites have an open API that allow you to connect your application to their site (Facebook, Twitter, Wikipedia, ...)
- Still such APIs often have drawbacks for network data collection:

Open APIs

- Most popular social web sites have an open API that allow you to connect your application to their site (Facebook, Twitter, Wikipedia, ...)
- Still such APIs often have drawbacks for network data collection:
  - Limited access rights (e.g. Facebook)

Open APIs

- Most popular social web sites have an open API that allow you to connect your application to their site (Facebook, Twitter, Wikipedia, ...)
- Still such APIs often have drawbacks for network data collection:
  - Limited access rights (e.g. Facebook)
  - Limited queries per time interval (e.g. Twitter) etc.

Outline

Networks are most often analyzed using graph theory:

Definition

A graph $\mathcal{G}$ is the pair $(\mathcal{N}, \mathcal{E})$ whereby $\mathcal{N}$ represents the set of verticles or nodes, and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ the set of edges connecting pairs from $\mathcal{N}$.

Networks are most often analyzed using graph theory:

Definition

A graph $\mathcal{G}$ is the pair $(\mathcal{N}, \mathcal{E})$ whereby $\mathcal{N}$ represents the set of verticles or nodes, and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ the set of edges connecting pairs from $\mathcal{N}$.

Definition

A directed graph or digraph $\mathcal{G}$ is the pair $(\mathcal{N}, \mathcal{A})$, whereby $\mathcal{N}$ represents the set of nodes, and $\mathcal{A} \subseteq \mathcal{N} \times \mathcal{N}$ the set of ordered pairs of elements from $\mathcal{N}$ that represent the set of graph arcs.

Networks are most often analyzed using graph theory:

Definition

A graph $\mathcal{G}$ is the pair $(\mathcal{N}, \mathcal{E})$ whereby $\mathcal{N}$ represents the set of verticles or nodes, and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ the set of edges connecting pairs from $\mathcal{N}$.

Definition

A directed graph or digraph $\mathcal{G}$ is the pair $(\mathcal{N}, \mathcal{A})$, whereby $\mathcal{N}$ represents the set of nodes, and $\mathcal{A} \subseteq \mathcal{N} \times \mathcal{N}$ the set of ordered pairs of elements from $\mathcal{N}$ that represent the set of graph arcs.

Definition

A valued or weighted graph $\mathcal{G}_{\mathcal{V}}$ is the triple $(\mathcal{N}, \mathcal{A}, \mathcal{V})$ whereby $\mathcal{N}$ represents the set of nodes or verticles, $\mathcal{A} \subseteq \mathcal{N} \times \mathcal{N}$ the set of pairs of elements from $\mathcal{N}$ that represent the set of graph arcs, and $\mathcal{V} : \mathcal{N} \to \mathbb{R}$ a function that attaches values or weights to nodes.

Definition

A bipartite graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ *is a graph for which set of nodes there exists an partition* $\mathcal{N} = \{X, Y\}$*, such that every arc has one end in* X*, a and the other in* Y*.*

Definition

A bipartite graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is a graph for which set of nodes there exists an partition $\mathcal{N} = \{X, Y\}$, such that every arc has one end in X, a and the other in Y.

Definition

A multigraph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is a graph in which $\mathcal{E}$ is a multiset, e.g. there can exist multiple edges between two nodes.

Matrix representation

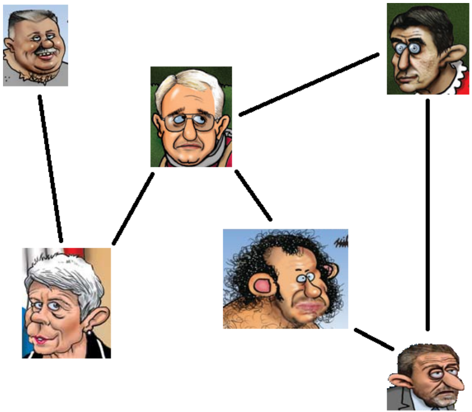
Definition

Let $\mathcal{G}$ be a graph defined with the set of nodes $\{n_1, n_2, ..., n_m\}$ and edges $\{e_1, e_2, ..., e_l\}$. For every i, j ($1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant m$) we define

$$a_{ij} = \begin{cases} 1, & \text{if there is an edge between nodes } n_i \text{ and } n_j \\ 0, & \text{otherwise} \end{cases}$$

Matrix $A = [a_{ij}]$ is then the adjacency matrix of graph $\mathcal{G}$. The matrix i symmetric since if there is an edge between nodes n_i and n_j then clearly there is also an edge between n_j and n_i. Thus $A = [a_{ij}] = [a_{ji}]$.

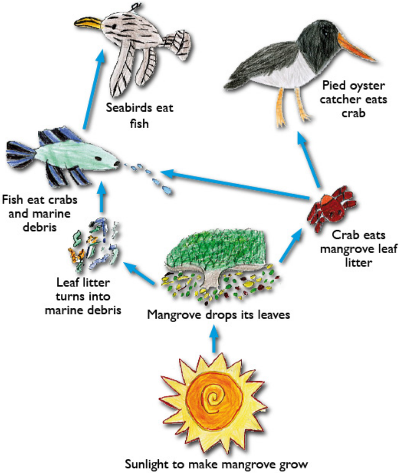
Simple graph



	0	1	0	0	0	1
	1	0	0	1	1	0
	0	0	0	1	0	0
	0	1	1	0	0	0
	0	1	0	0	0	1
	1	0	0	0	1	0

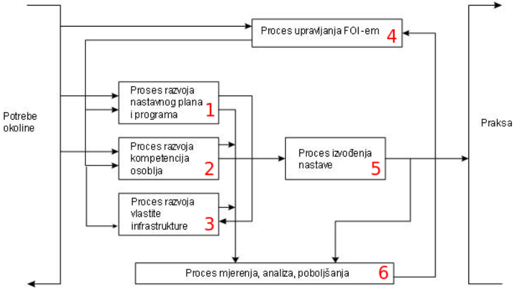
Slike by

Digraph



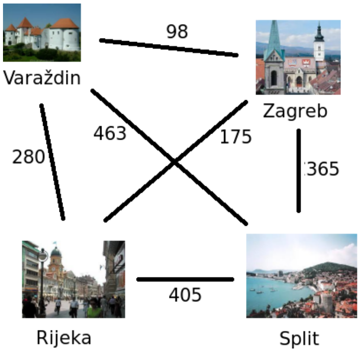
							
	0	0	0	0	0	0	0
	1	0	0	0	1	0	0
	0	0	0	0	0	0	0
	0	1	0	0	0	1	0
	0	0	1	0	0	0	0
	0	0	0	0	1	0	0
	0	0	0	1	0	0	0

Digraph



$$
\begin{array}{c|cccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
1 & 0 & 0 & 1 & 0 & 0 & 1 \\
2 & 0 & 0 & 0 & 0 & 1 & 1 \\
3 & 0 & 0 & 0 & 0 & 0 & 1 \\
4 & 1 & 1 & 1 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 0 & 0 & 1 \\
6 & 0 & 0 & 0 & 1 & 0 & 0
\end{array}
$$

Weighted graph

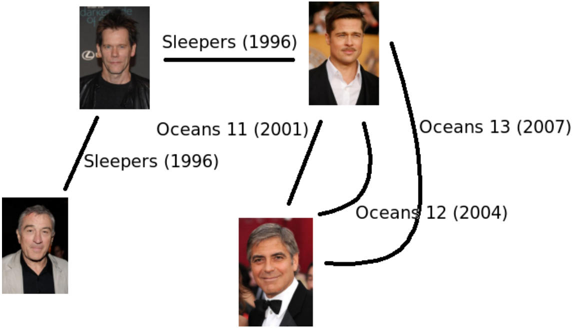


	Varaždin	Zagreb	Rijeka	Split
Varaždin	**0**	**98**	**280**	**463**
Zagreb	**98**	**0**	**175**	**365**
Rijeka	**280**	**175**	**0**	**405**
Split	**463**	**365**	**405**	**0**

Bipartite graph

Captain	1	1	0	0	0	0
Commander	0	1	1	0	0	0
Lt. Junior Grade	0	0	0	1	1	0
Lieutenant	0	0	0	1	1	0
Lt. Commander	0	0	0	1	1	1

Multigraph



0	1	0	3
1	0	1	0
0	1	0	0
3	0	0	0

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...
Examples:
$\mathbb{P}$ - process flow matrix ($p \times p$, p - number of processes)

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...
Examples:
$\mathbb{P}$ - process flow matrix ($p \times p$, p - number of processes)
$\mathbb{W}$ - co-worker matrix ($w \times w$, w - number of workers)

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...

Examples:

$\mathbb{P}$ - process flow matrix ($p \times p$, p - number of processes)

$\mathbb{W}$ - co-worker matrix ($w \times w$, w - number of workers)

$\mathbb{O}$ - process responsibility matrix ($p \times z$)

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...

Examples:

$\mathbb{P}$ - process flow matrix ($p \times p$, p - number of processes)

$\mathbb{W}$ - co-worker matrix ($w \times w$, w - number of workers)

$\mathbb{O}$ - process responsibility matrix ($p \times z$)

$\mathbb{WW}$ - co-workers of co-workers

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...

Examples:

$\mathbb{P}$ - process flow matrix ($p \times p$, p - number of processes)

$\mathbb{W}$ - co-worker matrix ($w \times w$, w - number of workers)

$\mathbb{O}$ - process responsibility matrix ($p \times z$)

$\mathbb{W}\mathbb{W}$ - co-workers of co-workers

$\mathbb{P}^T$ - inversed process flow

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...
Examples:
$\mathbb{P}$ - process flow matrix ($p \times p$, p - number of processes)
$\mathbb{W}$ - co-worker matrix ($w \times w$, w - number of workers)
$\mathbb{O}$ - process responsibility matrix ($p \times z$)
$\mathbb{W}\mathbb{W}$ - co-workers of co-workers
$\mathbb{P}^T$ - inversed process flow
$\mathbb{O}^T\mathbb{O}$ - matrix of workers who are responsible for the same processes

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...

Examples:

$\mathbb{P}$ - process flow matrix ($p \times p$, p - number of processes)

$\mathbb{W}$ - co-worker matrix ($w \times w$, w - number of workers)

$\mathbb{O}$ - process responsibility matrix ($p \times z$)

$\mathbb{W}\mathbb{W}$ - co-workers of co-workers

$\mathbb{P}^T$ - inversed process flow

$\mathbb{O}^T\mathbb{O}$ - matrix of workers who are responsible for the same processes

$\mathbb{P}\mathbb{O}$ - matrix of workers who are responsible for the forthcoming processes

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...

Examples:

$\mathbb{P}$ - process flow matrix ($p \times p$, p - number of processes)

$\mathbb{W}$ - co-worker matrix ($w \times w$, w - number of workers)

$\mathbb{O}$ - process responsibility matrix ($p \times z$)

$\mathbb{W}\mathbb{W}$ - co-workers of co-workers

$\mathbb{P}^T$ - inversed process flow

$\mathbb{O}^T\mathbb{O}$ - matrix of workers who are responsible for the same processes

$\mathbb{P}\mathbb{O}$ - matrix of workers who are responsible for the forthcoming processes

$\mathbb{P}^T\mathbb{O}$ - matrix of workers who are responsible for the previous processes

Matrix algebra and interpretation

Matrices can be transposed, multiplied ...
Examples:
$\mathbb{P}$ - process flow matrix ($p \times p$, p - number of processes)
$\mathbb{W}$ - co-worker matrix ($w \times w$, w - number of workers)
$\mathbb{O}$ - process responsibility matrix ($p \times z$)
$\mathbb{W}\mathbb{W}$ - co-workers of co-workers
$\mathbb{P}^T$ - inversed process flow
$\mathbb{O}^T\mathbb{O}$ - matrix of workers who are responsible for the same processes
$\mathbb{P}\mathbb{O}$ - matrix of workers who are responsible for the forthcoming processes
$\mathbb{P}^T\mathbb{O}$ - matrix of workers who are responsible for the previous processes
...

ACI model (Mika 2007)

- *A* - actors

ACI model (Mika 2007)

- *A* - actors
- *C* - concepts

ACI model (Mika 2007)

- *A* - actors
- *C* - concepts
- *I* - instances

ACI model (Mika 2007)

- A - actors
- C - concepts
- I - instances

- $T \subseteq A \times C \times I$ - *folksonomy*

ACI model (Mika 2007)

- A - actors
- C - concepts
- I - instances

- $T \subseteq A \times C \times I$ - *folksonomy*

A tripartite hypergraph which can be reduced into three bipartite graphs:

ACI model (Mika 2007)

- A - actors
- C - concepts
- I - instances

- $T \subseteq A \times C \times I$ - *folksonomy*

A tripartite hypergraph which can be reduced into three bipartite graphs:

- AC - network of actors and concepts

ACI model (Mika 2007)

- A - actors
- C - concepts
- I - instances

- $T \subseteq A \times C \times I$ - *folksonomy*

A tripartite hypergraph which can be reduced into three bipartite graphs:

- AC - network of actors and concepts
- AI - network of actors and instances

ACI model (Mika 2007)

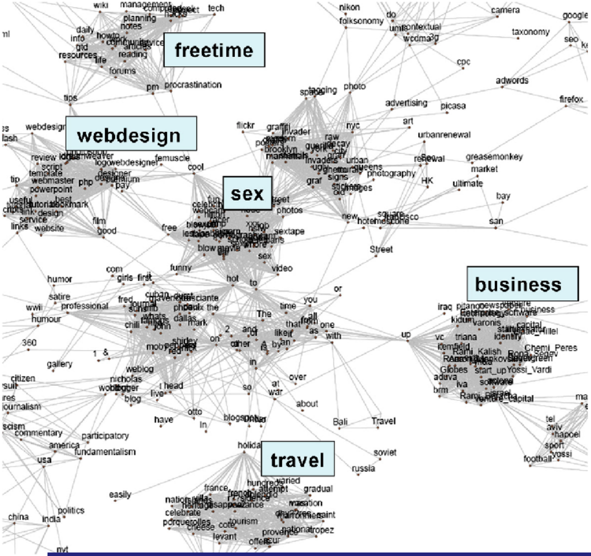
- A - actors
- C - concepts
- I - instances

- $T \subseteq A \times C \times I$ - *folksonomy*

A tripartite hypergraph which can be reduced into three bipartite graphs:

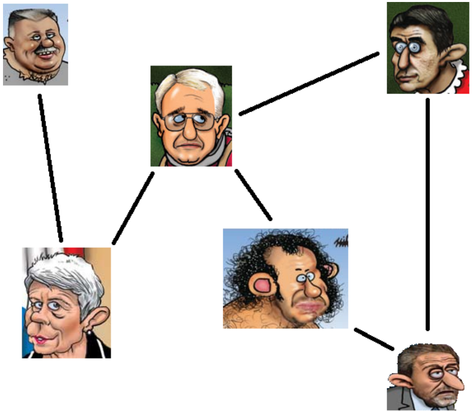
- AC - network of actors and concepts
- AI - network of actors and instances
- CI - network of concepts and instances

Delicious - analysis



delicious

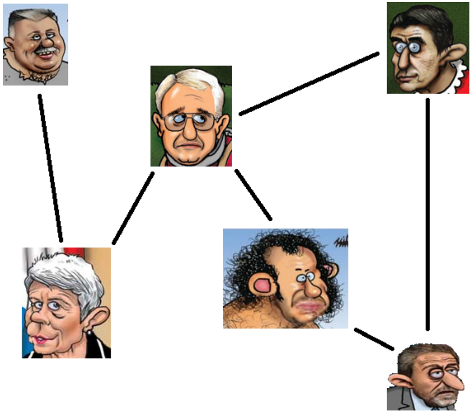
Problems with matrix representation



	0	1	0	0	0	1
	1	0	0	1	1	0
	0	0	0	1	0	0
	0	1	1	0	0	0
	0	1	0	0	0	1
	1	0	0	0	1	0

Slike by

Problems with matrix representation

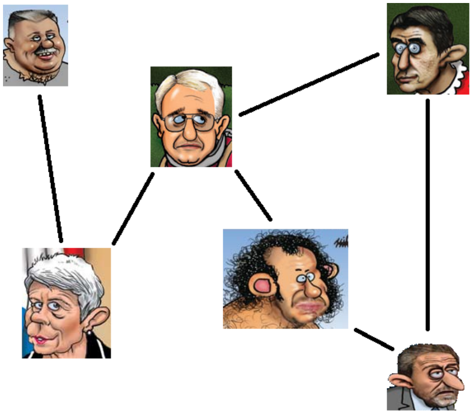


0	1	0	0	0	1
1	0	0	1	1	0
0	0	0	1	0	0
0	1	1	0	0	0
0	1	0	0	0	1
1	0	0	0	1	0

Slike by

- Only the “1”-s contain information

Sparse matrix



Slike by

Problems with sparse matrix

- If A is connected to B, then B is connected to A (undirected graphs)

Problems with sparse matrix

- If A is connected to B, then B is connected to A (undirected graphs)
- How to find friends of friends?

Problems with sparse matrix

- If A is connected to B, then B is connected to A (undirected graphs)
- How to find friends of friends?
- Is there a path from node A to node B? Which path is that?

Problems with sparse matrix

- If A is connected to B, then B is connected to A (undirected graphs)
- How to find friends of friends?
- Is there a path from node A to node B? Which path is that?
- How to compute the previously outlined graphs?

Problems with sparse matrix

- If *A* is connected to *B*, then *B* is connected to *A* (undirected graphs)
- How to find friends of friends?
- Is there a path from node *A* to node *B*? Which path is that?
- How to compute the previously outlined graphs?

Logic programming as a solution

- Deductive languages like Prolog, $\mathcal{F}$LORA-*2*

Logic programming as a solution

- Deductive languages like Prolog, $\mathcal{F}$LORA-*2*
- Advantages:
  - Declarative problem solving (“on a higher level”)
  - Very appropriate for expressing mathematical structures

Logic programming as a solution

- Deductive languages like Prolog, $\mathcal{F}$LORA-*2*
- Advantages:
  - Declarative problem solving ("on a higher level")
  - Very appropriate for expressing mathematical structures
- Disadvantages:
  - Combinatoric explosion (use heuristics and predicate tabling)

Examples

If *A* is connected to *B*, then *B* is connected to *A*

```
link( A, B ) :- link( B, A ).
```

Examples

Friend of a friend

```
ff( A, B ) :- link( A, C ), link( C, B ).
```

Examples

Is there a path from node *A* to node *B*?

```
path( A, B ) :- link( A, B ).
path( A, B ) :- link( A, C ), path( C, B ).
```

Examples

Which path is that?

```
whichpath( X, Y, [ X, Y ] ) :- link( X, Y ).
whichpath( X, Y, [ X, Z | T ] ) :-
      link( X, Z ),
      whichpath( Z, Y, [ Z | T ] ).
```

Outline

Project outline

- Croatian Scientific Bibliography (`http://bib.irb.hr`)

Project outline

- Croatian Scientific Bibliography (`http://bib.irb.hr`)
- Scientists enter data about their research activities by them selves which makes it a social application

Project outline

- Croatian Scientific Bibliography (`http://bib.irb.hr`)
- Scientists enter data about their research activities by them selves which makes it a social application
- Research hypotheses:

Project outline

- Croatian Scientific Bibliography (`http://bib.irb.hr`)
- Scientists enter data about their research activities by them selves which makes it a social application
- Research hypotheses:
  1. analyze two social systems (the Croatian scientific community and the Croatian public) and see how the most important research concepts (as viewed by the scientific community) are interpreted by the public (does the public understand most important concepts from Croatian science in the last decade?).

Project outline

- Croatian Scientific Bibliography (`http://bib.irb.hr`)
- Scientists enter data about their research activities by them selves which makes it a social application
- Research hypotheses:
  1. analyze two social systems (the Croatian scientific community and the Croatian public) and see how the most important research concepts (as viewed by the scientific community) are interpreted by the public (does the public understand most important concepts from Croatian science in the last decade?).
  2. analyze how various scientific fields are conceptually and socially interrelated (do scientists do interdisciplinary research together if they have conceptually connected fields?)

Methodology

- Social systems are represented by their “trails”: CROSBI as the Croatian scientific community, Croatian Wikipedia as the Croatian public
- One system can interpret concepts from another if the concept exists within its conceptual network (e.g. keyword or wiki page)

Data collection

- CROSBI data has been collected using Scrapy in November 2010 (a total of 285,234 entries has been collected)
- Wikipedia data has been collected using the Wikipedia API + Scrapy for parsing

Results

Bibliography

1. Adamic, L.: **Why networks are interesting to study**, University of Michigan, School of Information, `https://open.umich.edu/education/si/si508/fall2008`
2. Barratm M., Barthlemy, M., Vespignani, A.: **Dynamical Processes on Complex Networks**, Cambridge University Press, 2008.
3. Carley, K.M.: **Dynamic Network Analysis**, Summary of the NRC workshop on social network modeling and analysis, Committee on Human Factors, National Research Council, 133145, Eds: Breiger, R., Carley, K.M., and Pattison, P., 2003.
4. Divjak, B., Lovrenčić, A.: **Diskretna matematika s teorijom grafova**, TIVA & FOI, 2005
5. Krackhardt, David, and Carley, Kathleen M.: **A PCANS Model of Structure in Organization**, Proceedings of the 1998 International Symposium on Command and Control Research and Technology, Evidence Based Research, Vienna, VA, June 1998
6. Mika, Peter: **Ontologies are us: A unified model of social networks and semantics**, Web Semantics: Science, Services and Agents on the World Wide Web 5(1), volume 5, 515, March 2007
7. Newman, M., Barabsi, A.-L., Watts, D. j.: **The Structure and Dynamics of Networks**, Princeton University Press, 2006.
8. Various web sites (images, graphs)