Introduction to data networks analysis

Lovro Šubelj

lovro.subelj@fri.uni-lj.si

Data Technologies Laboratory
Faculty for Computer and Information science
University of Ljubljana

April 2010

Introduction
- Graphs & networks
- Network data model

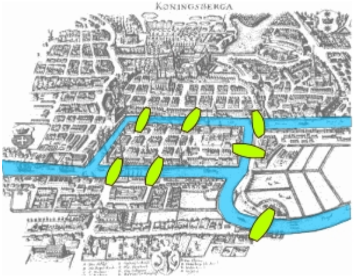
Real-world networks
- Types
- Examples
- Properties
- Complexity

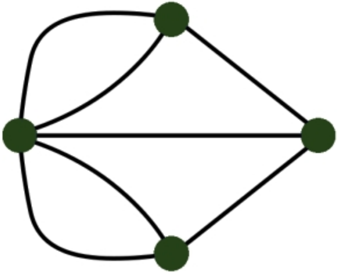
Network analysis
- PageRank
- HITS
- Infomap

Applications
- Fraud detection
- Telecommunications
- Biology
- WWW

Motivation

Seven bridges of Königsberg (Euler, 1735).



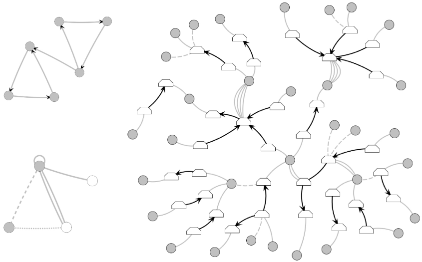


Graphs

Graph G is a discrete mathematical object that consists of:

- a set of nodes $V = \{v_1, v_2 \dots v_n\}$ and
- a set of edges among nodes $E \subseteq \{\{v_i, v_j\} |\ v_i, v_j \in V\}$.

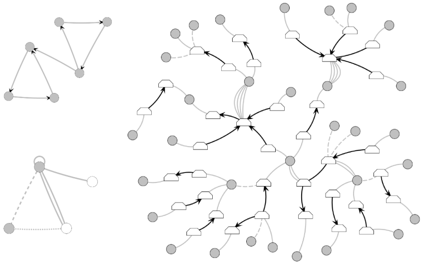
Furthermore, graph can be directed, labeled, weighted, multi-, hyper-, etc.



Networks

Networks are graphs with some additional information on the nodes and edges. Network N consists of:

- a set of nodes V and edges E,
- node features $F_V : V \to \Sigma_V$ and
- edge features $F_E : E \to \Sigma_E$.



Network data model

Why networks?

- Natural representation of many domains.
- Strong mathematical foundations.
- Useful whenever we are interested in relations among entities.

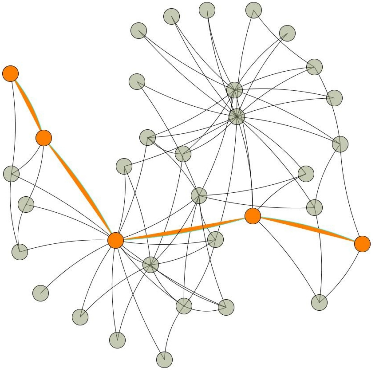
Network data model has to store:

- nodes, edges and features,
- their incidence.

Standard data models (e.g. relational, hierarchical, etc.) are **not appropriate** for storing networks.

Querying

Toy example: find the shortest path from node v_i to v_j in a simple graph G.



Querying, cont.

Toy example: find the shortest path from node v_i to v_j in a simple graph G.

Relational data model (e.g. *Oracle 11g*)

- Schema $Edge(\#\underline{node}1, \#\underline{node}2)$.
- Use Dijkstra's algorithm with v_i being the source node.
- Algorithm does $\mathcal{O}(|E|)$ queries over the database.

Network data model (e.g. *Oracle 11g Spatial*)

- Schemas *Node*, *Link*, *Path*, etc.
- Use `manager.shortestPath(`G`,`v_i`,`v_j`)` (Java API).
- Algorithm does a single (?) query over the database!

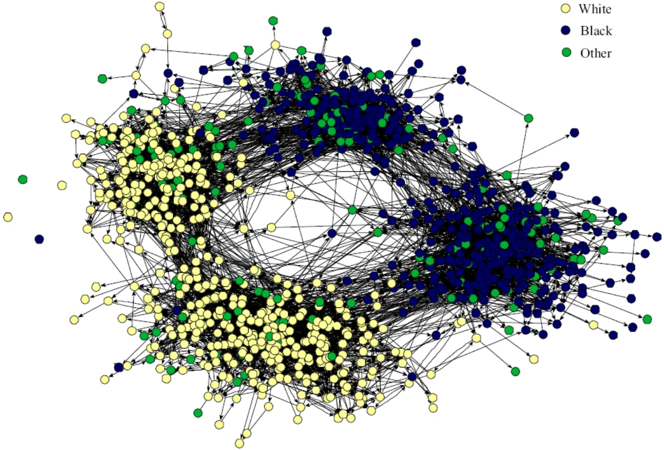
Types of networks

Real-world networks are of different types:

- social (e.g. network of Twitter, friendship network),
- information (e.g. WWW, citation networks),
- biological (e.g. gene regulatory networks, food webs),
- technological (e.g. Internet, power grid),
- etc.

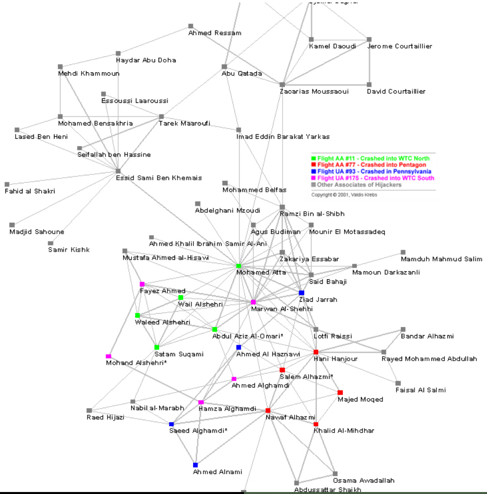
Examples - social networks

Friendship network of children in U.S. school.



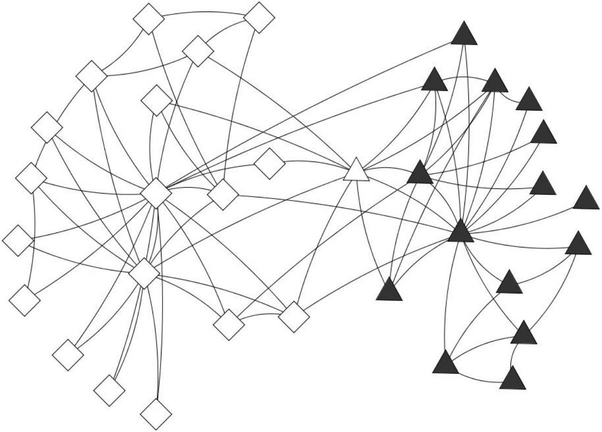
Examples - social networks

Krebs's terrorist network of 9-11.



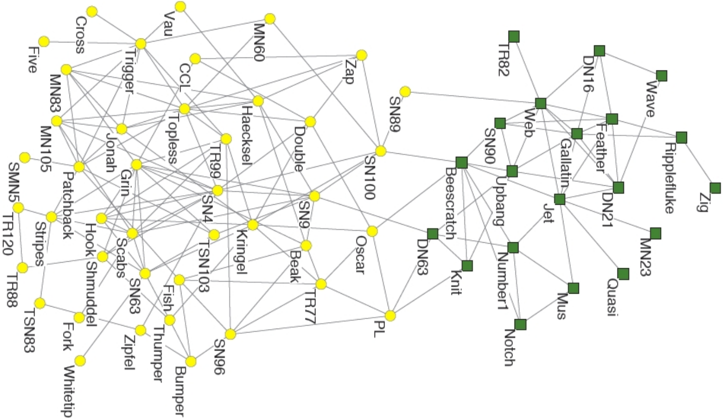
Examples - social networks

Zachary's karate club network.



Examples - social networks

Lusseau's bottlenose dolphins network.

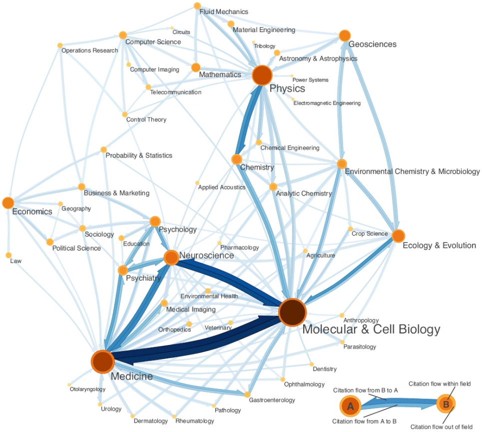


Examples - social networks

Social network of Jesus.

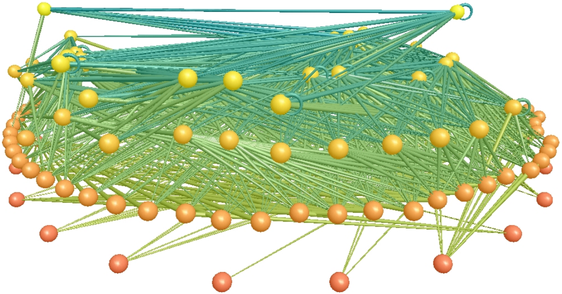
Examples - information networks

Map of science.



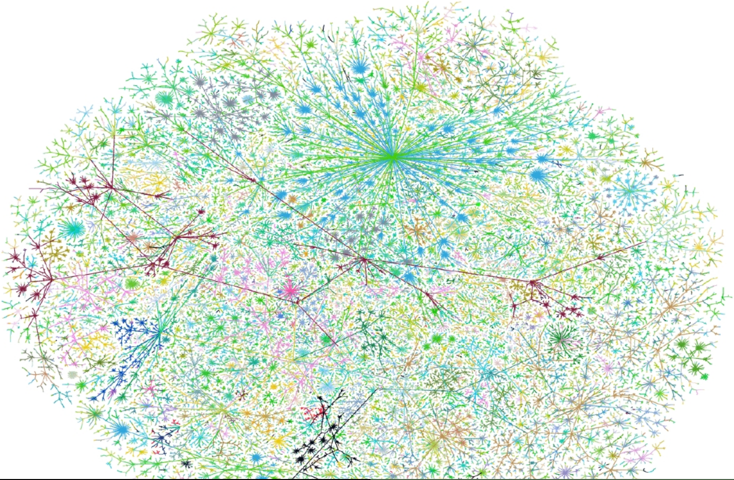
Examples - biological networks

Food web of predator-prey interactions.



Examples - technological networks

Internet at the level of "autonomous systems".



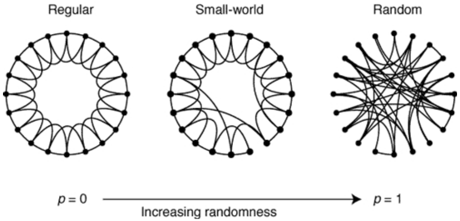
Properties of networks

Properties of many real-world networks:

- *small-world effect*,
- *power law*,
- transitivity,
- community structure,
- network resilience,
- network navigation,
- etc.

Small-world effect

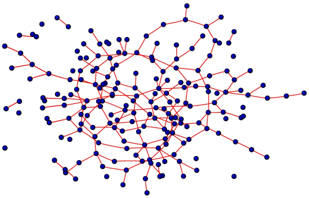
- Famous Milgram's passing letters experiment.
- *Small-world effect* (*6 degrees of separation*): the average distance among nodes is small, even when the network is very large.

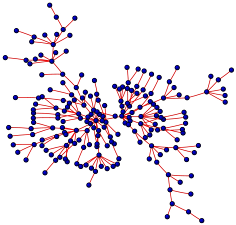


- The effect is actually mathematically obvious.
- Applications for dynamics of processes on networks.

Power law

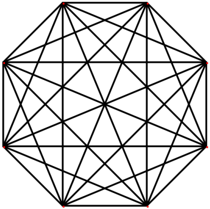
- *Power law* (*scale-free networks*): the degree distribution of nodes is power law in the tail, i.e., there exist nodes with extremely high degree.
- Possible explanation is *preferential attachment* (i.e. *rich-get-richer*).





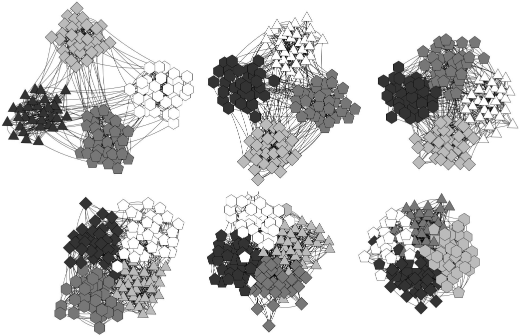
Transitivity

- Transitivity (clustering): "friend of a friend is often also a friend".
- Transitivity means presence of a high number of triangles.



Community structure

- Community structure: existence of groups of nodes (i.e. communities) with many edges within communities and only a few edges between communities.



Complexity

Networks can be classified as follows:

- **small** tens of nodes,
- **medium** several hundreds of nodes,
- **large** several thousands or millions of nodes,
- **huge** several millions of nodes and more.

- Even medium networks are commonly hard to visualize.
- Large networks are those that can still be stored in computer's memory, whereas huge networks cannot – a problem as nodes cannot be analyzed independently!

Many of problems over networks are NP-hard.

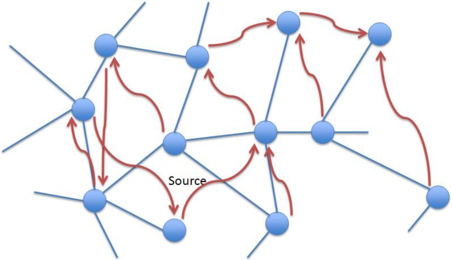
Network analysis

Network analysis (from AI perspective) is a broad field:

- (social) network analysis (SNA),
- link analysis (LA),
- graph-based data mining (GBDM),
- (statistical) relational learning (SRL),
- community detection,
- etc.

PageRank (Brin and Page, 1998)

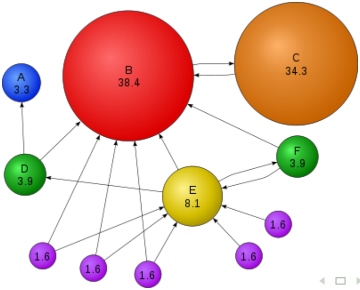
- A link analysis algorithm that exploits hyperlinks among web pages.
- Result is a ranking of web pages (due to their importance).
- Main hypothesis: random walker on a (WWW) network would spend most of the time on "important" web pages, and less on "unimportant" ones.



PageRank, cont.

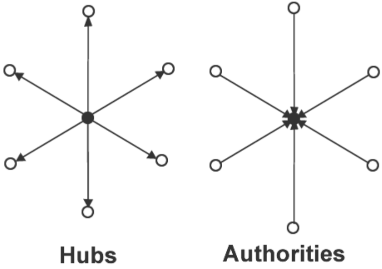
- Result of a random walk corresponds to the first eigenvector of the adjacency matrix of the network.
- *PageRank* algorithm (d is a damping factor, e.g., 0.85):

$$PR(v_i) = \frac{1-d}{|V|} + d \sum_{v_j \in \mathcal{N}(v_i)} \frac{PR(v_j)}{deg^+(v_j)}$$



HITS (Kleinberg, 1999)

- A link analysis algorithm that exploits hyperlinks among web pages (*Hyperlink-Induced Topic Search*).
- Result is a ranking of web pages (due to their *authority* and *hub* importance).
- Main hypothesis: page is a good hub if it links to many good authorities (and vice versa).

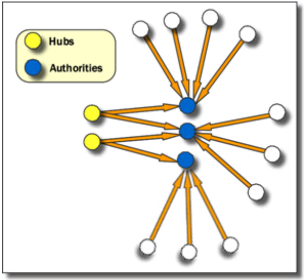


HITS, cont.

- *HITS* algorithm:

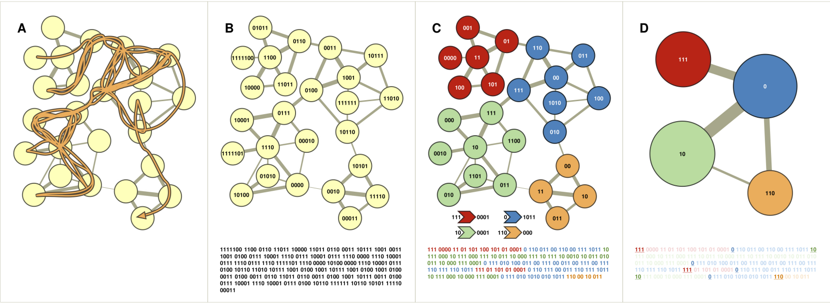
$$a(v_i) = \sum_{(v_j, v_i) \in E} h(v_j)$$

$$h(v_i) = \sum_{(v_i, v_j) \in E} a(v_j)$$



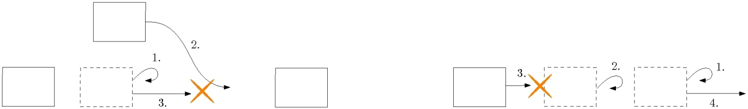
Infomap (Rosvall and Bergstrom, 2008)

- A community detection algorithm that again uses random walks on the network and Huffman coding.
- Result is a clustering of the nodes of the graph.



Automobile insurance fraud

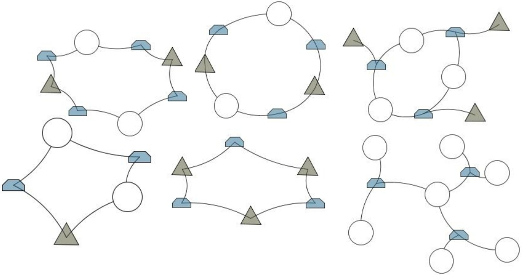
- Focus on groups of collaborating fraudsters staging traffic "accidents". Groups consist of drivers, chiropractics, police officers, lawyers, insurance workers, etc.
- Common characteristics of such accidents (e.g. at night, no children, no narcotics, etc.).
- Standard schemes for staging traffic accidents like *Drive Down* and *Swoop and Squat*.



- Such accidents are very expensive and dangerous.

Suspicious patterns detection

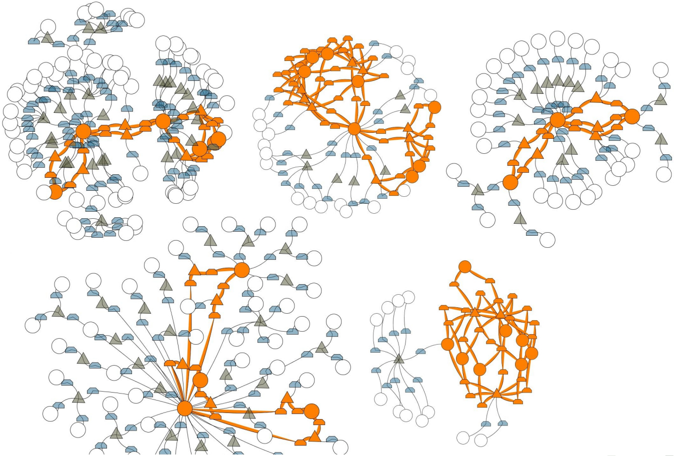
- There are many suspicious patterns:



- Patterns can be detected using simple subgraph isomorphism.

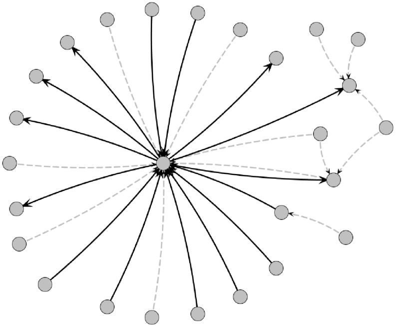
Suspicious patterns detection, cont.

Results of suspicious patterns detection.



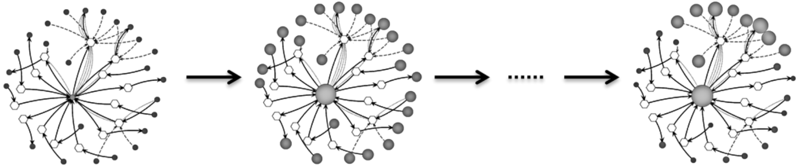
Suspicious components detection

- Detection of suspicious components using their common (topological) characteristics (e.g. cycles, diameter, central nodes, etc.).
- *PRIDIT* analysis of indicators of suspicious components.



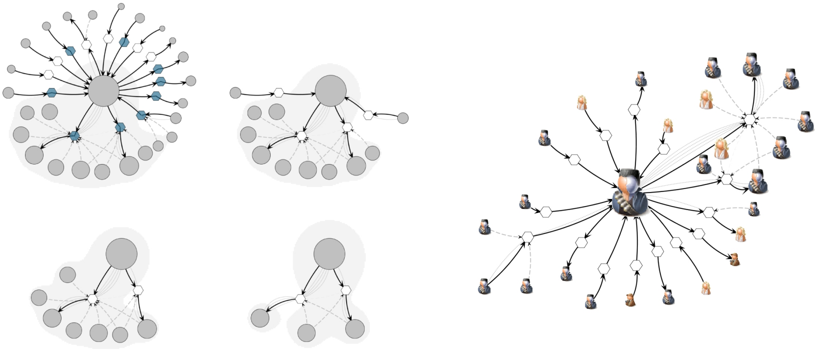
Suspicious entities detection

- Main hypothesis: suspicion of every entity is well defined with the suspicions of its related entities (participant-accident).
- Suspicion of every entity is inferred from the network and propagated onward using *PageRank*-like algorithm (i.e. *IAA* algorithm).
- Iterative assessment reduces locality of the approach.



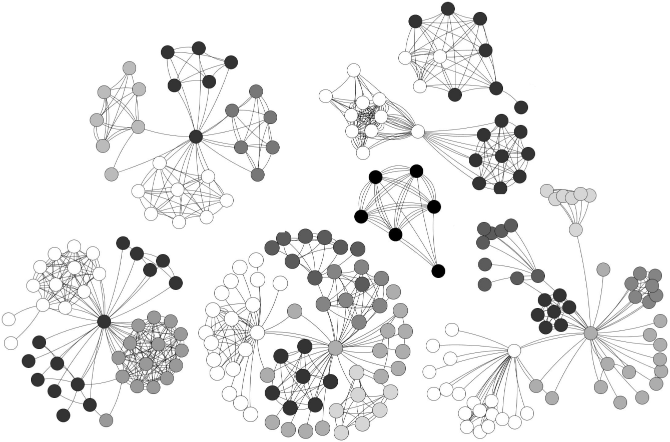
Suspicious entities detection, cont.

Results of suspicious entities detection.



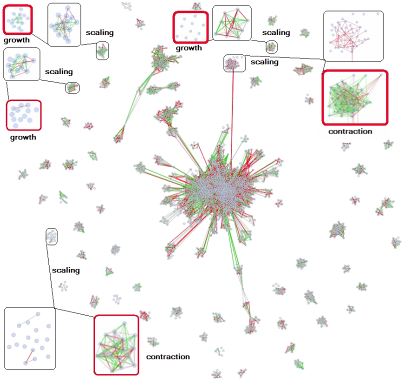
Suspicious communities detection

Results of “suspicious” communities detection.



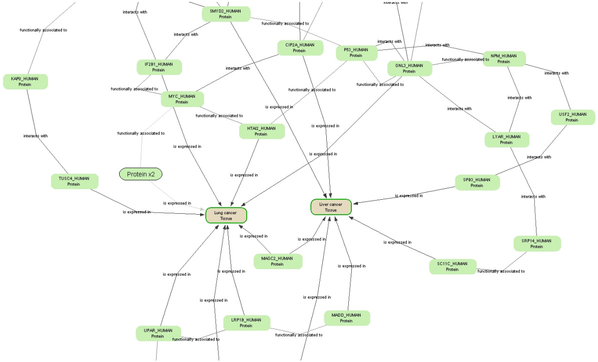
TeleComVis project (Beijing University)

- Visual analysis of telecom communities in massive call graphs.
- Main objective is to adjust market strategies to different social communities.



Biomine project (University of Helsinki)

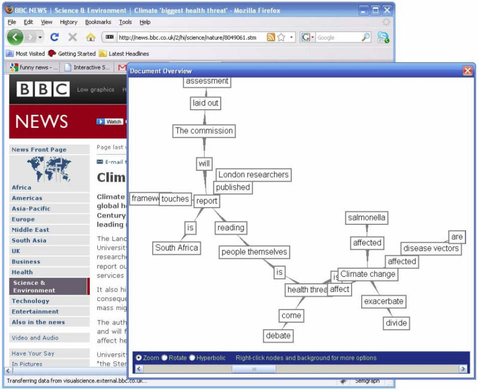
- Knowledge discovery from public biological databases using networks.
- Currently their database contains $\approx 10^6$ nodes ($\approx 6 \times 10^6$ edges) representing proteins, genes, articles, ontologies, etc.



WWW

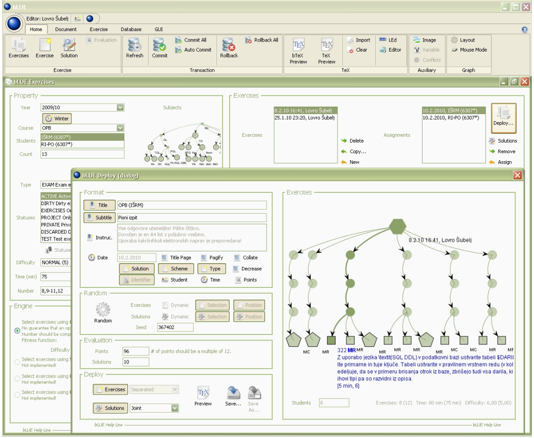
SemGraph project (Jožef Stefan Institute)

- Enhanced web page visualization using semantic graphs and text mining.
- Works as a Firefox plug-in.



bLUE project

- Automatic generation of exams from the database of exercises.
- Exams and exercises visualized with networks.



Questions?