



Introduction and Objectives

Social Network Analysis and Mining (SNAM) includes different techniques from sociology, social sciences, mathematics, statistics and computer science.

Objectives

- Analysis of the structure of a social network
- Analysis of large sub-networks and connected components
- Discovering nodes of particular interest
- Identifying communities within the network

Advantages

- Large scale studies, impossible before, are feasible
- Data can be automatically acquired
- A huge amount of information is accessible online
- Data could be acquired at different granularity level

Limits

- Problems related to large scale data mining issues
- Computational and algorithmic challenges
- Bias of data should be investigated

Web Data Extraction

WDE Systems Software platform for the extraction, in an automatic and intelligent fashion, of data from Web pages, under the form of static and/or dynamic contents, in order to store them in a database (or other structured data sources) and make them available for other applications.

Wrapper An algorithmic procedure which aims to the extraction of unstructured information from a data source (such as a Web page) and transform it in a structured format.

Wrapper maintenance Web pages changes, dynamically, not only in contents but also in the structure, without any forewarning. Web wrappers could stop working properly, because of radical or even small changes in that structure.

Automatic Wrapper Adaptation A novel smart approach to automatize the work of maintenance has been proposed.

Clustered Tree Matching

HTML Web pages are represented as trees, whose nodes contains elements displayed in the page.
XPath A standard language defined to identify elements within a Web page. Wrappers implements the XPath logic.

Clustered Tree Matching key aspects (Ferrara, 2011)

- Inspired by Simple Tree Matching (STM) (Selkow, 1977)
- Assigns weights to evaluate importance of matches
- Different behavior considering leaves or middle-level nodes
- Introduces a degree of accuracy
- Identify clusters of similar sub-trees

Algorithm 1 ClusteredTreeMatching(T' , T'')

```
1: if  $T'$  has the same label of  $T''$  then
2:    $m \leftarrow d(T')$ 
3:    $n \leftarrow d(T'')$ 
4:   for  $i = 0$  to  $m$  do
5:      $M[i][0] \leftarrow 0$ 
6:   for  $j = 0$  to  $n$  do
7:      $M[0][j] \leftarrow 0$ 
8:   for all  $i$  such that  $1 \leq i \leq m$  do
9:     for all  $j$  such that  $1 \leq j \leq n$  do
10:       $M[i][j] \leftarrow \max(M[i][j-1], M[i-1][j], M[i-1][j-1] + W[i][j])$  where  $W[i][j] = \text{ClusteredTreeMatching}(T'(i-1), T''(j-1))$ 
11:   if  $m > 0$  AND  $n > 0$  then
12:     return  $M[m][n] * 1 / \max(t(T'), t(T''))$ 
13:   else
14:     return  $M[m][n] + 1 / \max(t(T'), t(T''))$ 
15:   else
16:     return 0
```

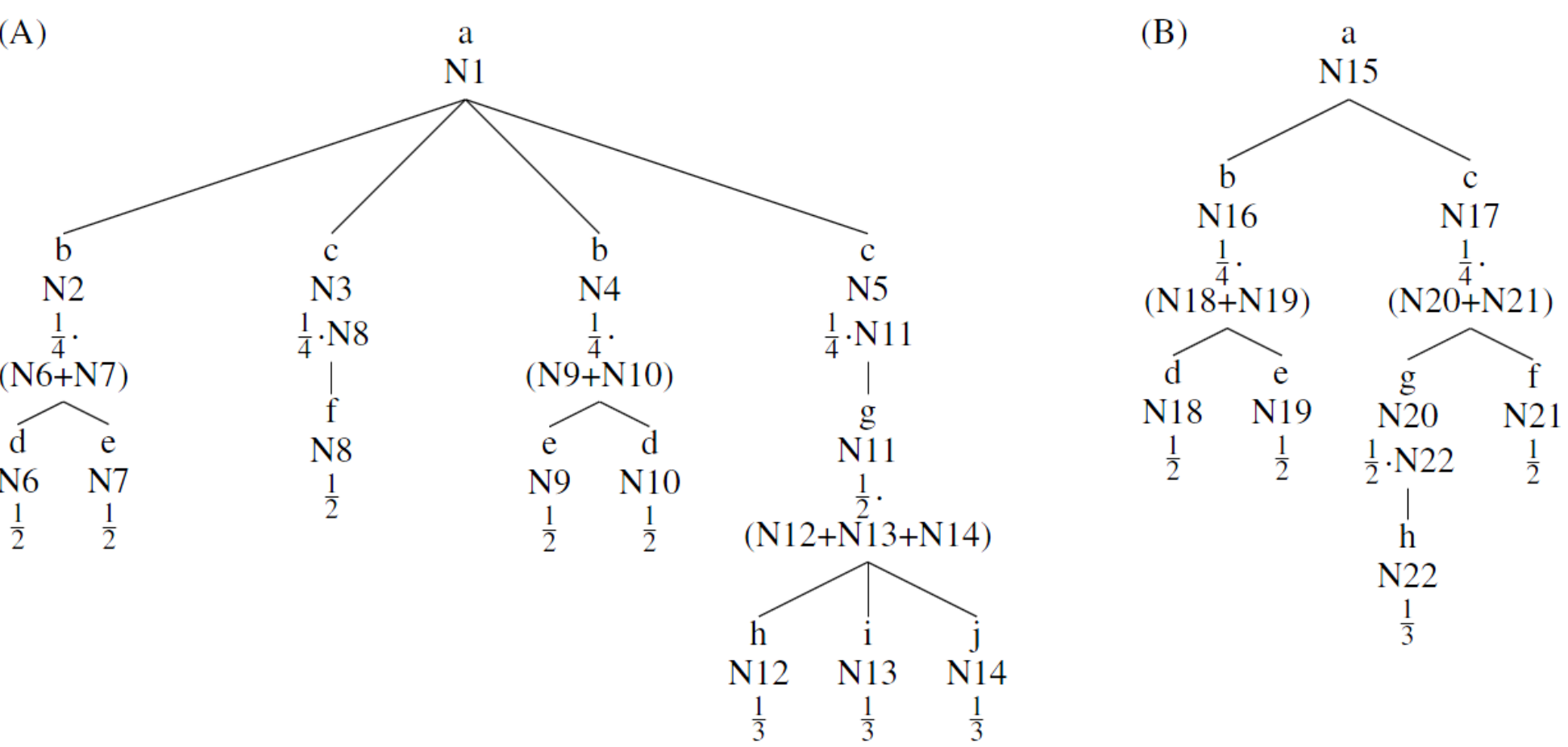


Figure: A and B are two similar trees. CTM assigns weights to matching nodes. Node **f** in **A** has weight $\frac{1}{2}$ because in **B** it appears in a sub-tree with two children. Node **h** in **B** has weight $\frac{1}{3}$ for the same reason.

Pros vs Cons

Advantages

- CTM produces an intrinsic measure of similarity (STM calculates only the number of different elements)
- It is possible to introduce a degree of accuracy establishing a minimum similarity threshold, required to match two clusters
- The more the structure of compared trees is complex and similar, the better the similarity calculated by CTM is!

Limits CTM/STM do not produce perfect results when nodes are added/removed in middle-levels.

Further considerations Pure text can be matched using classic string matching techniques, e.g. Jaro-Winkler, bigrams, etc.

Example of Wrapper Adaptation with CTM

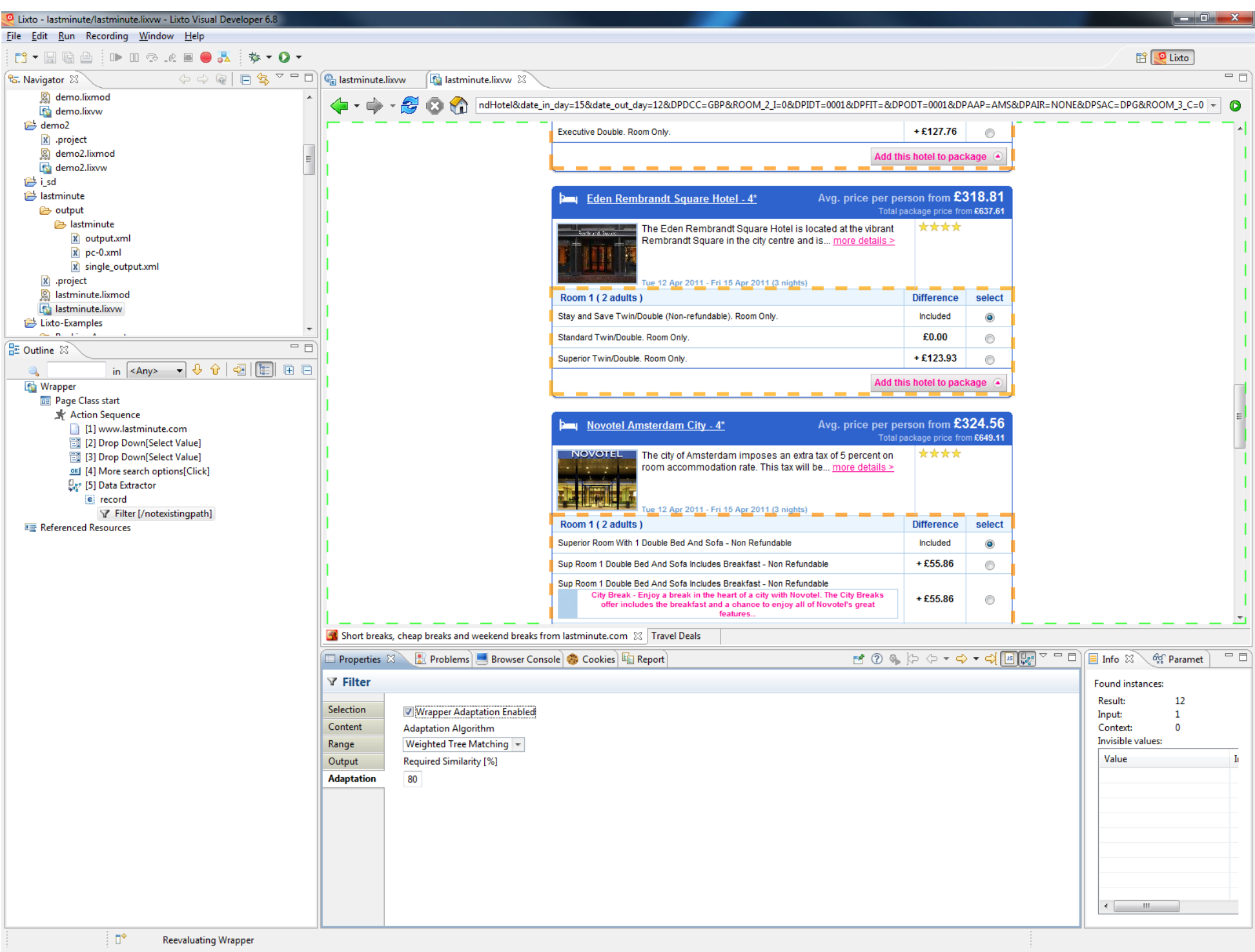


Figure: An example of automatic adaptation to modifications. Elements matched by the original wrapper are identified even with structural differences, by applying the CTM algorithm.

WDE Platform Architecture

Intelligent Agent It's a platform (software + architecture) which could autonomously take smart decisions to achieve a goal.

- Each Web wrapper is implemented as an Agent
- Several Agents populate the same environment
- If a Wrapper fails, it adapts itself to changes
- Results are collected in a transparent way w.r.t. users

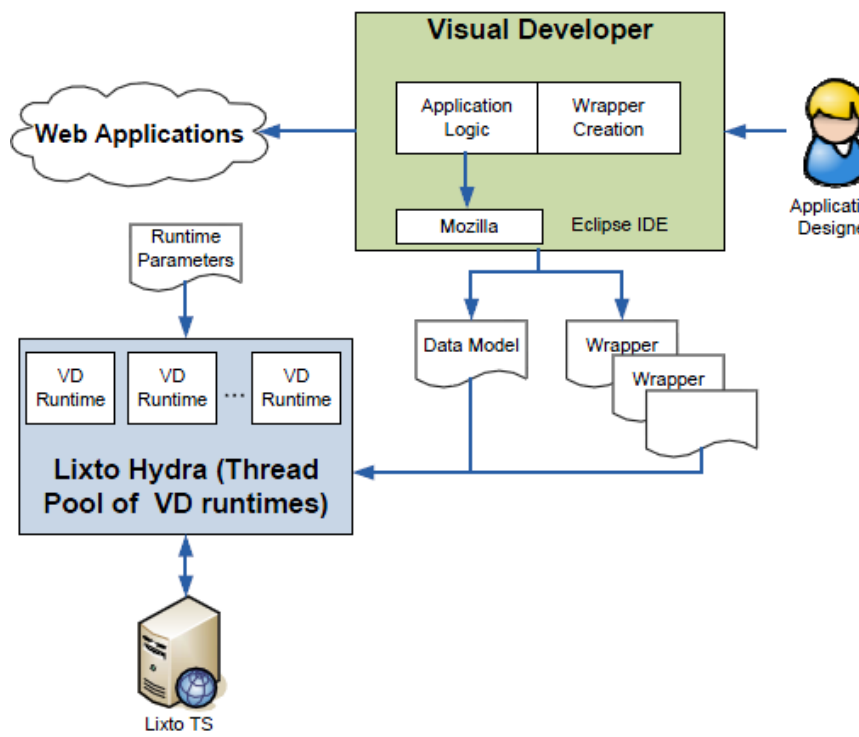


Figure: Web Data Mining platform architecture

BFS Sampling

BFS (breadth-first search): starting from a seed, a graph is visited exploring all the neighbors in order of discovering.

Pros

- Optimal solution for unweighted and/or undirected graphs (such as Facebook and other OSNs)
- Intuitive implementation

Cons Resulting samples are biased towards high degree nodes in incomplete visits.

Challenge Obtaining a sub-graph of the Facebook network which preserves properties of the complete graph.

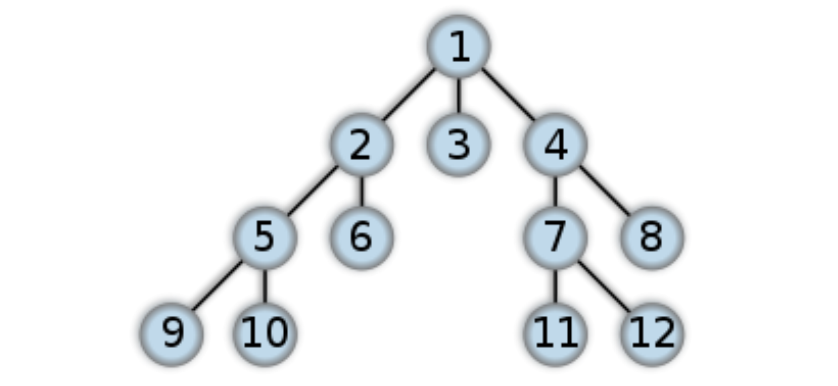


Figure: BFS (3rd sub-level)

- 1 seed
- 2-4 friends
- 5-8 friends of friends
- 9-12 friends of friends of friends

Uniform Sampling

Uniform (rejection sampling): a list of random nodes to be visited is generated.

Pros

- Independent w.r.t. the structural distribution of friendship ties
- Produces unbiased results
- Simple and efficient implementation

Cons Resulting graph has disconnected components.

Challenge Acquiring a uniform sub-graph with a huge connected component.

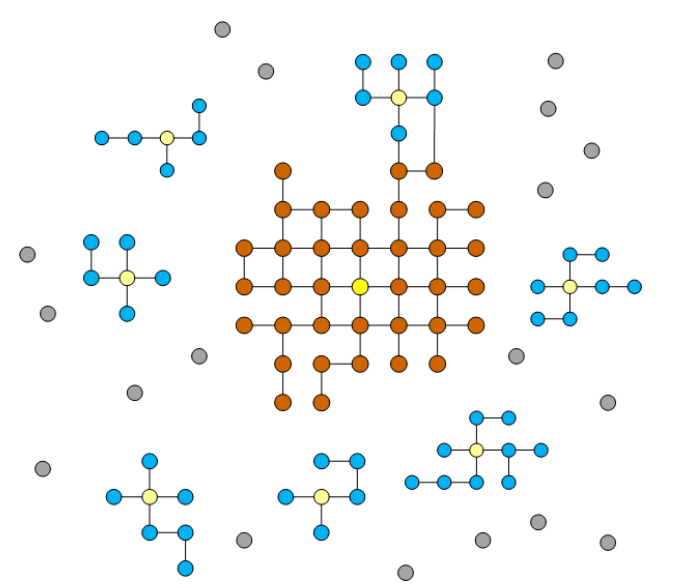


Figure: Uniform sampling

FBI Agent (Facebook Intelligent Agent)



Figure: FBI Agent Mining Architecture: Java application - Embedded Firefox browser / HTTP Interface → Facebook



Figure: FBI Agent Logic of Crawling: Diagram of the process of data extraction from Facebook

Crawling Facebook with CTM

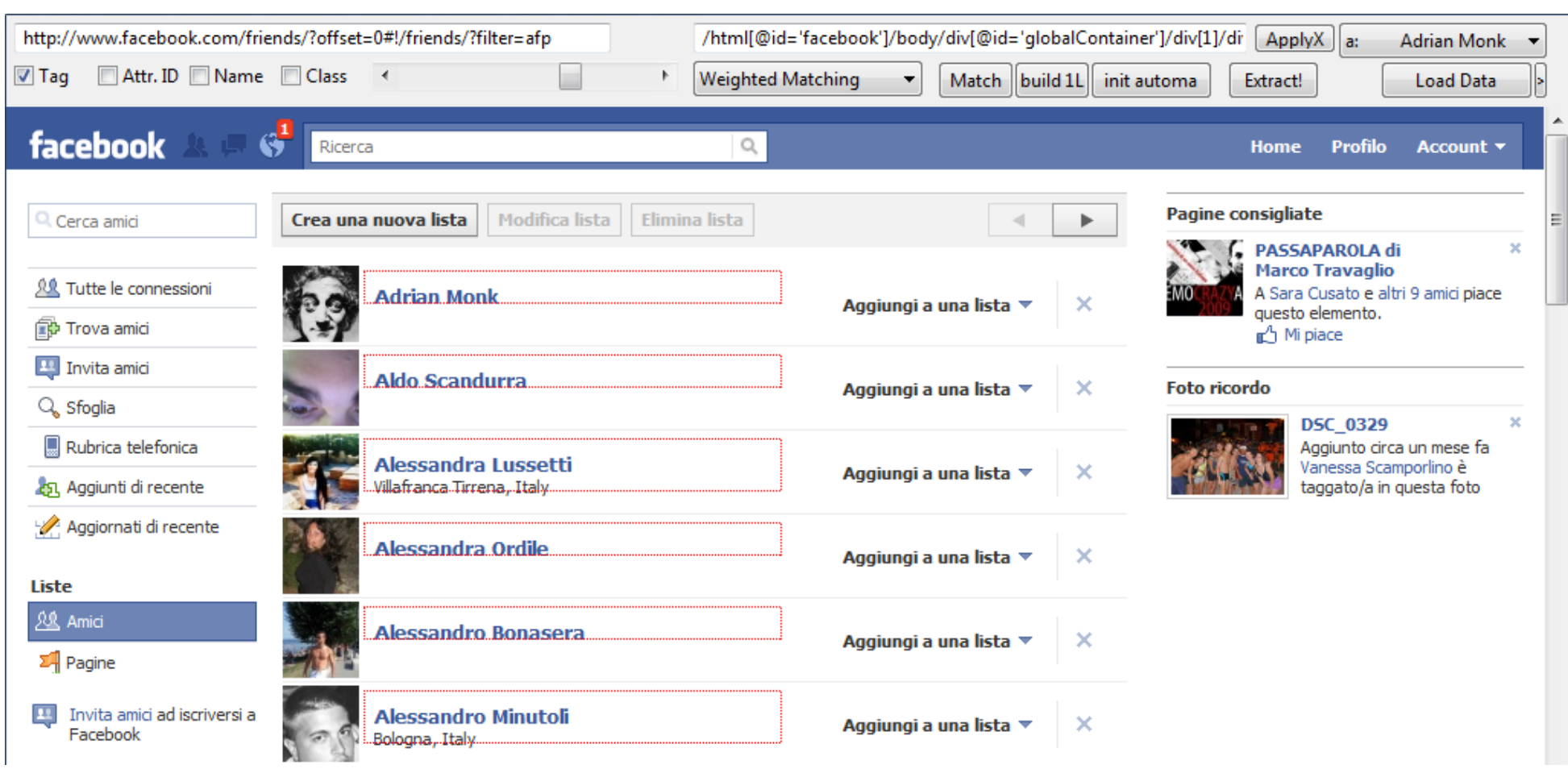


Figure: FBI Agent i) visits the page containing the friendlist, ii) generates a Wrapper to extract Name and ID of each friend, iii) insert into the graph these data, and iv) processes next profile.

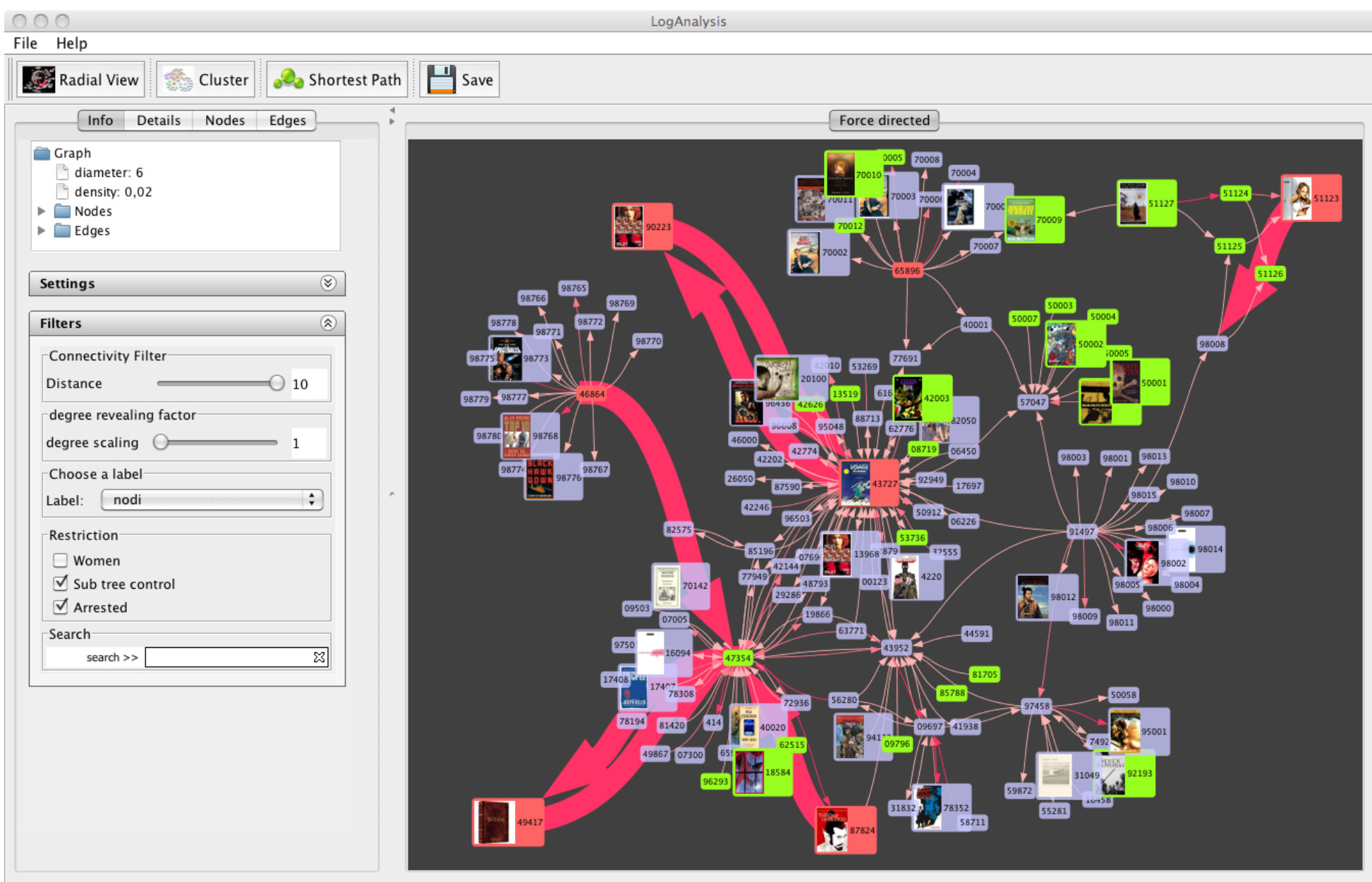
Dataset: BFS and Uniform (August 2010)

	N. Visited users	N. Discovered users	N. edges
BFS	63.4K	8.21M	12.58M
Uni	48.1K	7.69M	7.84M

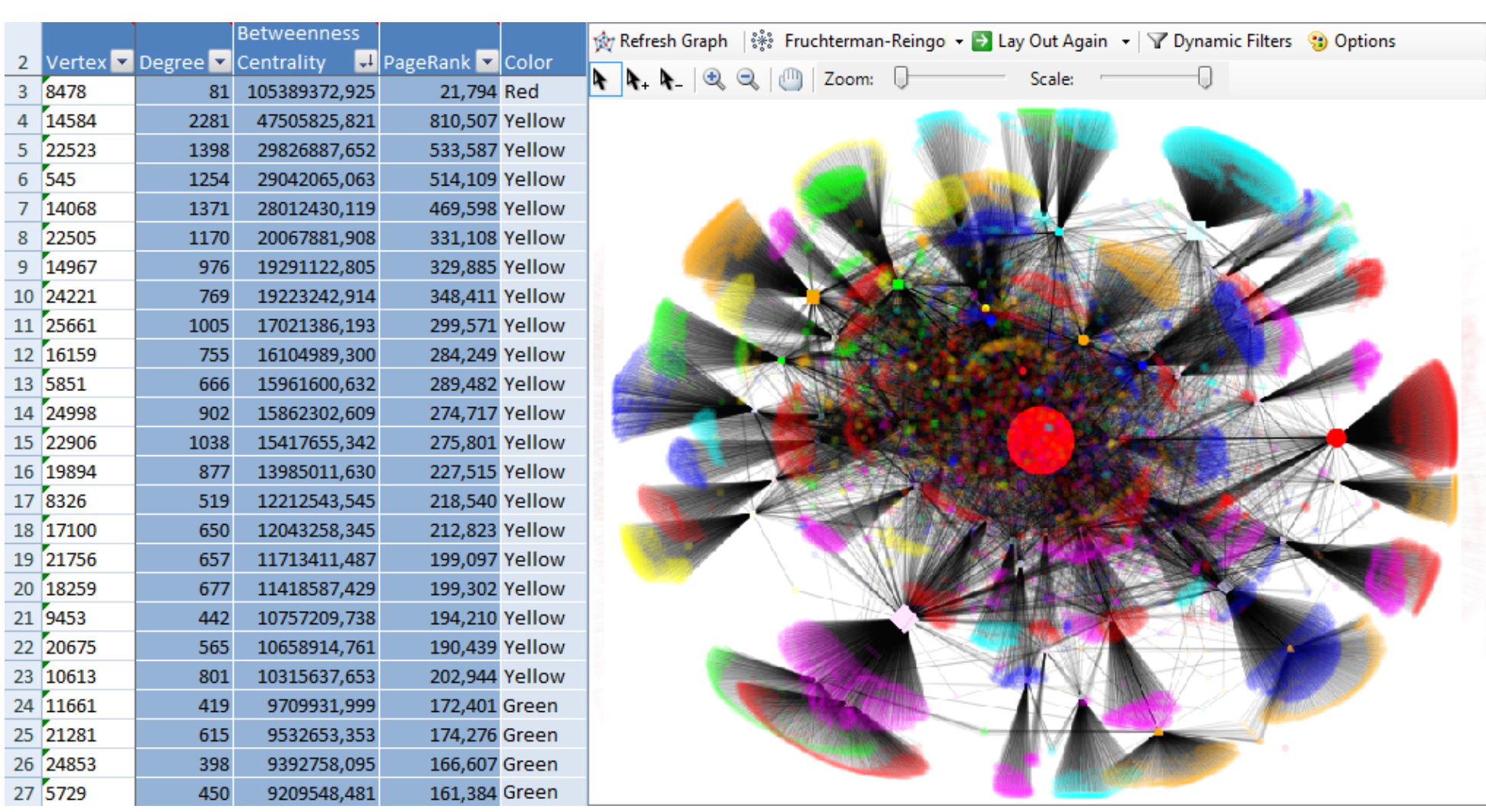
Avg. deg.	Eigenvec.	Diameter	Clustering	Coverage	Density
396.8	68.93	8.75	0.0789	98.98%	0.626%
326.0	23.63	16.32	0.0471	94.96%	0.678 %

Remarks Our data contain the same information as if we would acquire all the friendship relations among all the inhabitants of a middle-size town (e.g., 100k people).

Facebook Network Graph: Visual Results

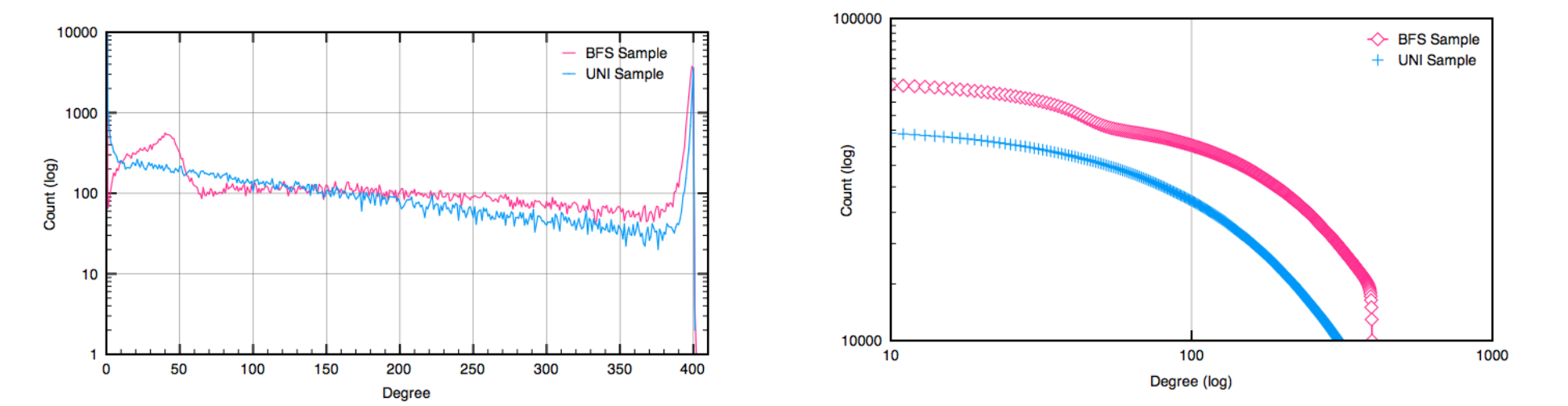


Betweenness Centrality Results



Node Degree Distribution

- Power-law distribution, $P(k) \sim k^{-\gamma}$, $\gamma \leq 3$
- This means the existence of a relatively small number of users highly connected each other
- This distribution could be represented by a Complementary Cumulative Distribution Function



References

- Ferrara E., Baumgartner R. **Automatic Wrapper Adaptation by Tree Edit Distance Matching**. In: *Combinations of Intelligent Methods and Applications* – Springer 2011.
- De Meo P., Ferrara E., Fiumara G. **Finding similar users in Facebook**. In: *Social Networking and Community Behavior Modeling: Qualitative and Quantitative Measures* – IGI Global 2011.
- Catanese S., De Meo P., Ferrara E., Fiumara G., Provetti A. **Crawling Facebook for social network analysis purposes**. In: *Int. Conf. on Web Intelligence, Mining and Semantics* – 2011.
- Ferrara E., Baumgartner R. **Design of automatically adaptable web wrappers**. In: *3rd International Conference on Agents and Artificial Intelligence* – 2011.
- Catanese S., De Meo P., Ferrara E., Fiumara G. **Analyzing the Facebook friendship graph**. In: *1st International Workshop on Mining the Future Internet* – 2010.