The 5th EuroSys Doctoral Workshop (EuroDW 2011)
Salzburg, Austria, Sunday 10 April 2011

Mining and Analyzing Online Social Networks

Emilio Ferrara
eferrara@unime.it

Advisor: Prof. Giacomo Fiumara

PhD School in Mathematics and Computer Science
Department of Mathematics
University of Messina

Outline

Outline

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.

Automatic Wrapper Adaptation A novel smart approach to make wrappers adaptive to structural changes 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.

Key aspects (Ferrara, 2011)

- Inspired by Simple Tree Matching (STM) [a]
- 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

[a] Tree to tree editing problem, Selkow, 1977

Algorithm 1 ClusteredTreeMatching(T', T'')

if T' has the same label of T'' **then**
  $m \leftarrow d(T')$
  $n \leftarrow d(T'')$
  for $i = 0$ to m **do**
    $M[i][0] \leftarrow 0$;
  for $j = 0$ to n **do**
    $M[0][j] \leftarrow 0$;
  for all i such that $1 \leq i \leq m$ **do**
    for all j such that $1 \leq j \leq n$ **do**
      $M[i][j] \leftarrow \text{Max}(M[i][j-1],\ M[i-1][j],\ M[i-1][j-1] + W[i][j])$ where $W[i][j] = \text{ClusteredTreeMatching}(T'(i-1), T''(j-1))$
  if $m > 0$ AND $n > 0$ **then**
    return M[m][n] * 1 / Max($t(T')$, $t(T'')$)
  else
    return M[m][n] + 1 / Max($t(T')$, $t(T'')$)
else
  return 0

Tree Matching Algorithm: Example (I)

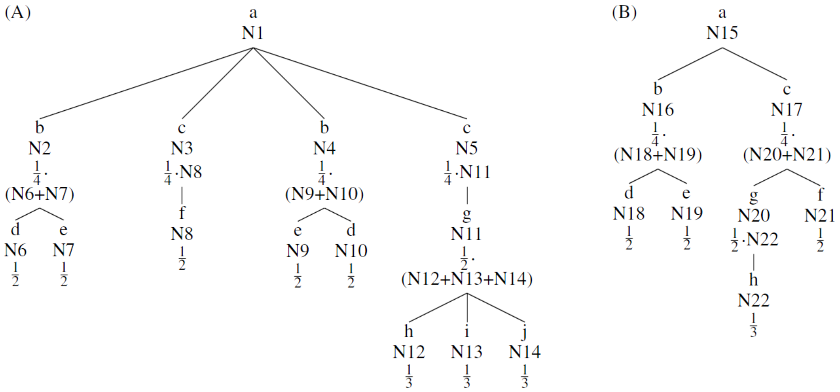


Figure: *A* and *B* are two similar trees. CTM assigns weights to matching nodes. Node *f* in *A* has weight $\frac{1}{2}$ because in *B* it appears in a sub-tree with two children. Node *h* in *B* has weight $\frac{1}{3}$ for the same reason.

Tree Matching Algorithm: Example (II)

W	N18	N19
N6	$\frac{1}{2}$	0
N7	0	$\frac{1}{2}$

M	0	N18	N18-19
0	0	0	0
N6	0	$\frac{1}{2}$	$\frac{1}{2}$
N6-7	0	$\frac{1}{2}$	1

W	N18	N19
N9	0	$\frac{1}{2}$
N10	$\frac{1}{2}$	0

M	0	N18	N18-19
0	0	0	0
N9	0	0	$\frac{1}{2}$
N9-10	0	$\frac{1}{2}$	$\frac{1}{2}$

W	N8
N20	0
N21	$\frac{1}{2}$

M	0	N8
0	0	0
N20	0	0
N20-21	0	$\frac{1}{2}$

W	N12	N13	N14
N22	$\frac{1}{3}$	0	0

M	0	N12	N12-13	N12-14
0	0	0	0	0
N22	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

W	N11
N20	$\frac{1}{6}$
N21	0

M	0	N11
0	0	0
N20	0	$\frac{1}{6}$
N20-21	0	$\frac{1}{6}$

W	N2	N3	N4	N5
N16	$\frac{1}{4}$	0	$\frac{1}{8}$	0
N17	0	$\frac{1}{8}$	0	$\frac{1}{24}$

M	0	N2	N2-3	N2-4	N2-5
0	0	0	0	0	0
N16	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
N16-17	0	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{3}{8}$

Figure: Couples of matrices *W* and *M*, step-by-step. CTM solves the similarity problem in 6 steps. Final result of similarity: $\frac{3}{8}$. Grey cells identify similar clusters between the two trees.

Outline

Auto-adaptive Web Wrappers: Requirements

For the implementation we identified

- Requirements:
  - The representation of the structure of the Web page [1]
  - If the original wrapper fails, it analyzes the tree structure of the page, identifying modifications
  - Once identified differences, the wrappers automatically adapts itself to the new structure
- Comparable elements:
  - Nodes: represent HTML elements, identified by HTML tags
  - Attributes: also attributes of nodes can be additionally compared

[1] using the syntax of tree-grams (tree-grammar) to simplify the representation

Auto-adaptive Web Wrappers: Example

Figure: An example of automatic adaptation of modifications. In the upper part of the screenshot the original page structure is shown. In the lower part, the new version. Modifications have been brought both to page structure and its contents. Elements matched by the original wrapper are even identified in the modified page, by applying the automatic adaptation policy.

Agent of Web data extraction

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

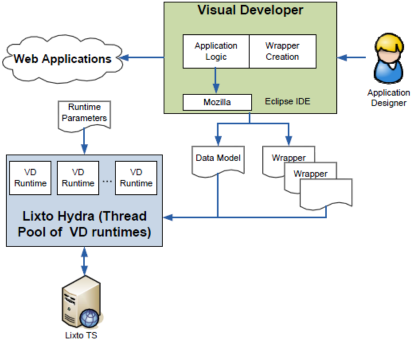


Figure: Web Data Mining platform architecture

Outline

Social Networks: Taxonomy

Social Networks (SN) A social network is a social structure made up of individuals (or organizations) connected each other, by (possibly) different social ties, such as friendship, kinship, shared interest, knowledge etc.

Social Network Analysis Analysis of social networks, (i.e., studying, modeling and measuring), could be conducted by using the formalism of graph theory.
Theory and models adopted for the study of SNs are part of the so called Social Network Analysis.

Several types of network exist: collaborations, communication, friendship, etc. Our study focuses on Online Social Networks:

- Social communities: **Facebook**, MySpace, etc.
- Sharing contents: YouTube, Flickr, etc.

Social Networks: Examples of OSN

Figure: An example of Online Social Network

Analysis of Online Social Networks: Motivations

Q: Is it possible to model social networks?

A: Analysis of characteristics and properties of OSNs graphs

Open problems

- Improving algorithms:
  - For visiting large graphs (e.g., BFS, Uniform, etc.)
  - To efficiently store and represent data (matrix decomposition, etc.)
  - Efficient and meaningful visualization of large graphs
  - Optimization of metrics calculation (e.g., All-Pairs-Shortest-Path, Betweenness Centrality, etc.)
- Investigation of the scalability
- Considering similitudes between OSNs and real social networks

Background and Related Work

Milgram The Small World problem (1970)

Zachary The first model of a real SN (1980)

Kleinberg Algorithmic perspective of SNs (2000)

Barabasi, Newman, et al. 2000+ focus on OSNs

- Large scale data mining from OSNs
- Visualization of large graphs
- Dynamics and evolution of OSNs
- SNA Metrics calculation
- Clustering, community structure, etc.

Remarks SNA is a “young” branch, born from the context of social sciences and moved towards mathematics and computer sciences in the last years.

Outline

Mining the Facebook graph: Breadth-first search

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.

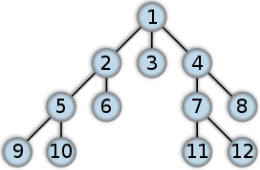


Figure: BFS (3rd sub-level)

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

Mining the Facebook graph: 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.

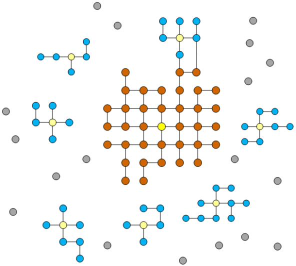


Figure: Uniform sampling

Mining the Facebook graph: How the Agent works

Initialization:

- Authentication on FB
- Selection of an example friendlist
- Generation of the wrapper for the automatic extraction

Execution:

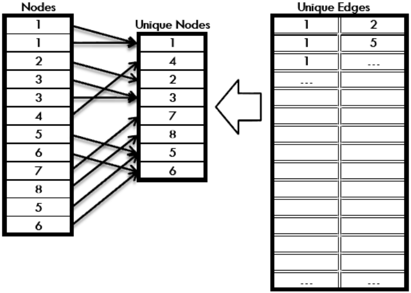
- Generation of a FIFO queue of profiles to be visited
- For each profile in queue:
  - Visit the friendlist page:
    - Extract friends (nodes) and relationships (edges)
    - (eventually) Put new friends in the FIFO queue
  - Cycle the process

Figure: Diagram of the process of data extraction from Facebook

Mining the Facebook graph: Data cleaning

Data cleaning $O(n)$ (optimal time)

1. Remove duplicates using hash tables
2. Delete parallel edges
3. Anonymize



Structuring data

Final data are stored as GraphML. It is a standard XML format for representing graph. It contains a description nodes within the graph and edges connecting them.

```
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
  http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">

  <!-- prefuse GraphML Writer | Wed Jul 07 11:14:43 CEST 2010 -->
  <key id="name" for="node" attr.name="name" attr.type="string"/>
  <key id="id" for="node" attr.name="id" attr.type="string"/>

  <graph edgedefault="undirected">
    <!-- nodes -->
    <node id="0">
      <data key="id">1659682073</data>
    </node>
    <node id="1">
      <data key="id">1370006884</data>
    </node>
    ...
    <!-- edges -->
    <edge id="0" source="1" target="2">
    </edge>
    ...
  </graph>
</graphml>
```

Mining the Facebook graph: Agent execution

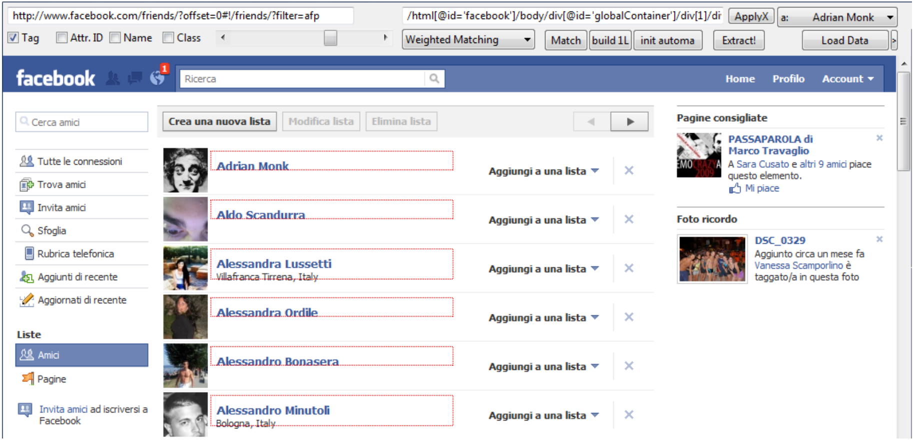
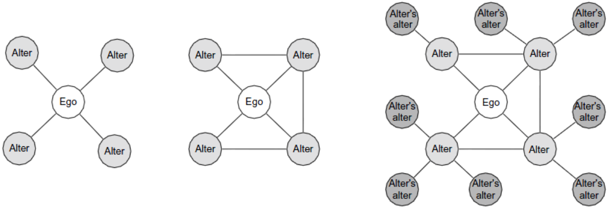


Figure: Agent i) visits the page containing the friendlist, ii) generates a Wrapper to extract Name and ID of each friend, iii) insert into the graph these data, and iv) proceeds with the next profile in the list.

Outline

Network Analysis Metrics: Characteristics of the Facebook graph

- **Ego-centric** network: the term *ego* denotes a user connected to others (alter)
- **Unweighted**, **undirected** network:
  - Degree 1.0
  - Degree 1.5
  - Degree 2.0



Remarks The graph shows a natural clustering effect over principal areas of the life of a user: friends, colleagues, family, etc.

Network Analysis Metrics: Measures

Perer and Shneiderman [2] provided a summary of useful metrics:

Overall metrics no. of nodes, edges, density, diameter, etc.

Centrality measures degree, betweenness and closeness centrality

Nodes in pairs plotting degree vs. betweenness

Cohesive sub-groups discovering communities, community structure

	N. Visited users	N. Discovered users	N. edges	Avg. deg.	Eigenvectors	Diameter	Clustering	Coverage	Density
BFS	63.4K	8.21M	12.58M	396.8	68.93	8.75	0.0789	98.98%	0.626%
Uni	48.1K	7.69M	7.84M	326.0	23.63	16.32	0.0471	94.96%	0.678 %

Table: Dataset: BFS and Uniform (acquired during August 2010)

[2] Balancing systematic and flexible exploration of social networks, 2006

Social Network Analysis Aspects: Visualization of data

Remarks Our data contains 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).

Visualization of Social Networks Providing a meaningful graphical representation of a large network in order to have greater insights on the structure, is a big challenge, both algorithmic and computational.

Problems

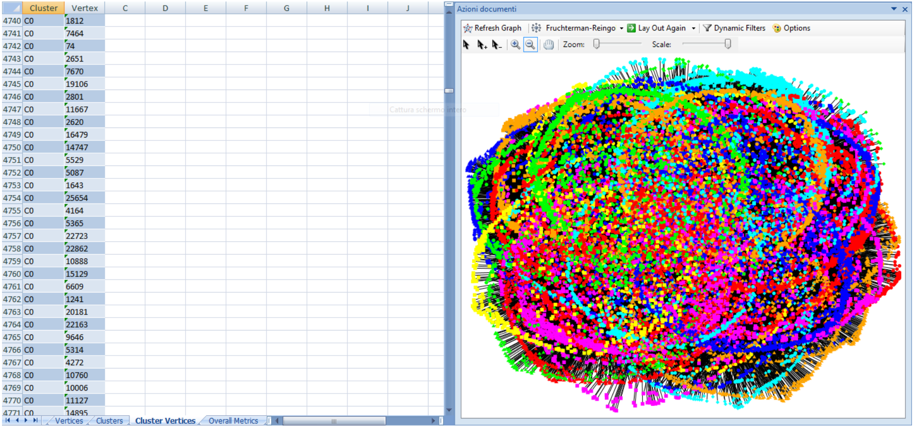
- The more the complexity of the network increases, the greater its illegibility is.
- Operations such as interaction on nodes and edges, filtering and manual positioning are required.

A: Our group [3] developed LogAnalysis, a powerful visual tool to analyze social network structures.

[3] A visual tool for forensic analysis of mobile phone traffic, Catanese & Fiumara, 2010

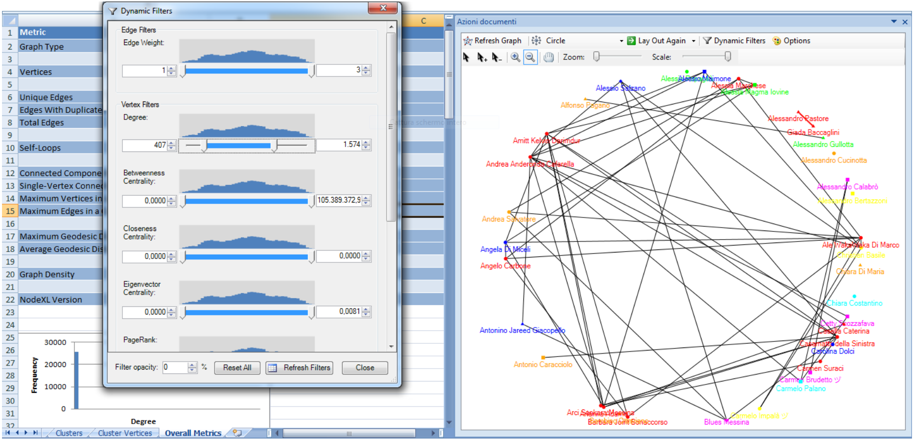
Facebook Network Graph: Visual results

NodeXL: Unfiltered graph (Dataset: 25K nodes sub-graph)



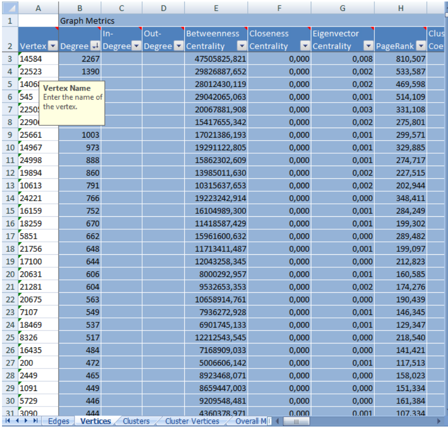
Facebook Network Graph: Visual results

NodeXL: Filtered graph (Dataset: 25K nodes sub-graph)

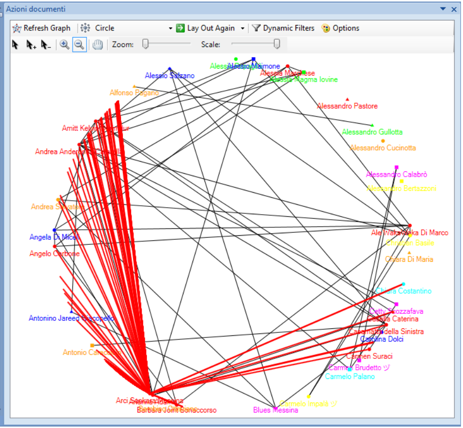


Facebook Network Graph: Visual results

NodeXL: Filtered results (Dataset: 25K nodes sub-graph)

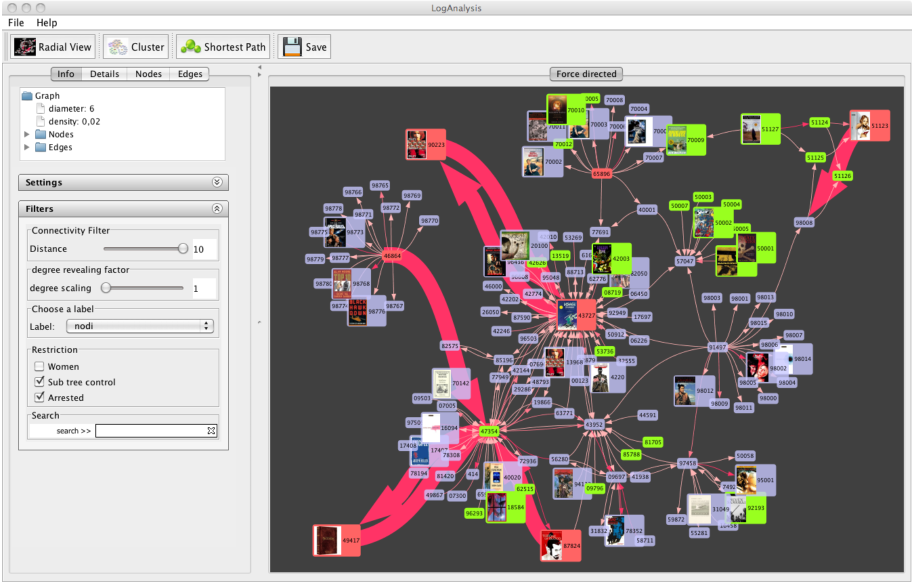


	A	B	C	D	E	F	G	H	
1		Graph Metrics							
2	Vertex	Degree	In-Degree	Out-Degree	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	PageRank	Clus Coe
3	14584	2267			47505825,821	0,000	0,008	810,507	
4	22523	1390			29826887,652	0,000	0,002	533,587	
5	1406				28012430,119	0,000	0,002	469,598	
6	545				29042065,063	0,000	0,001	514,109	
7	2250				20067881,908	0,000	0,003	331,108	
8	2290				15417655,342	0,000	0,002	275,801	
9	25661	1003			17021386,193	0,000	0,001	299,571	
10	14967	973			19291122,805	0,000	0,001	329,885	
11	24998	888			15862302,609	0,000	0,001	274,717	
12	19894	860			13985011,630	0,000	0,002	227,515	
13	10613	791			10315637,653	0,000	0,002	202,944	
14	24221	766			19223242,914	0,000	0,000	348,411	
15	16159	752			16104989,300	0,000	0,001	284,249	
16	18259	670			11418587,429	0,000	0,001	199,302	
17	5851	662			15961600,632	0,000	0,000	289,482	
18	21756	648			11713411,487	0,000	0,001	199,097	
19	17100	644			12043258,345	0,000	0,000	212,823	
20	20631	606			8000292,957	0,000	0,001	160,585	
21	21281	604			9532653,353	0,000	0,002	174,276	
22	20675	563			10658914,761	0,000	0,000	190,439	
23	7107	549			7936272,928	0,000	0,001	146,345	
24	18469	537			6901745,133	0,000	0,001	129,347	
25	8326	517			12212543,545	0,000	0,000	218,540	
26	16435	484			7168909,033	0,000	0,000	141,421	
27	200	472			5006606,142	0,000	0,001	117,513	
28	2449	465			8923468,071	0,000	0,000	158,023	
29	1091	449			8659447,003	0,000	0,000	151,334	
30	5729	446			9209548,481	0,000	0,000	161,384	
31	3090	444			4360378,971	0,000	0,001	107,334	



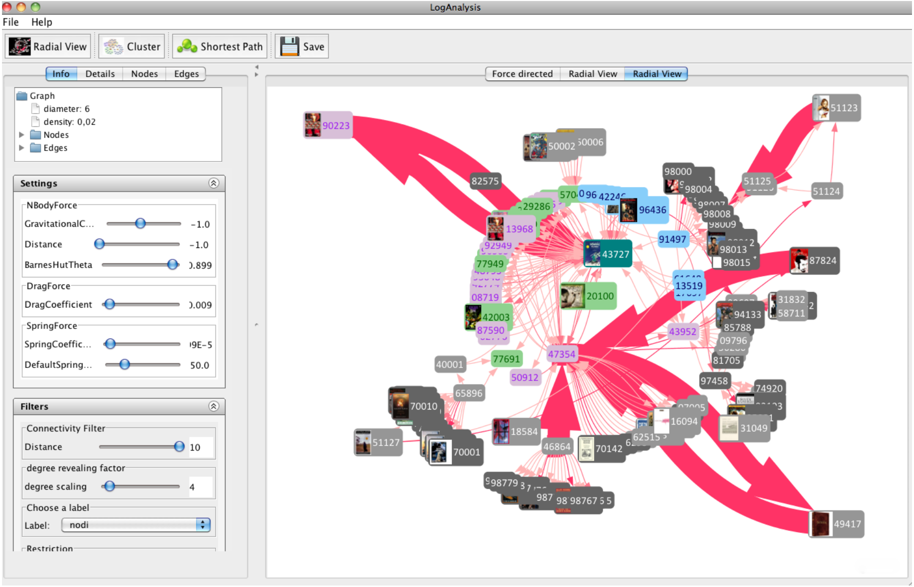
Facebook Network Graph: Visual results

LogAnalysis: Force Directed graph (Dataset: 25K nodes sub-graph)



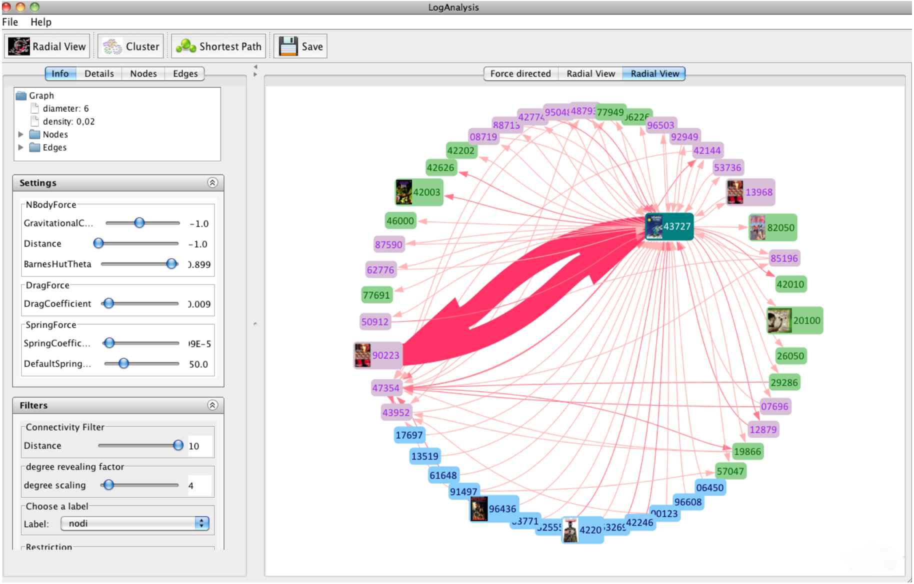
Facebook Network Graph: Visual results

LogAnalysis: Clustering (Dataset: 25K nodes sub-graph)



Facebook Network Graph: Visual results

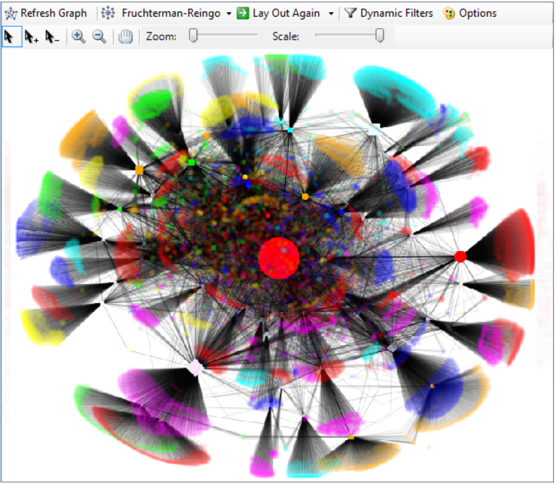
LogAnalysis: Radial view (2.0 degree)



Betweenness Centrality results

Top 25 Nodes ordered w.r.t. BC (Dataset: 25K nodes sub-graph)

2	Vertex	Degree	Betweenness Centrality	PageRank	Color
3	8478	81	105389372,925	21,794	Red
4	14584	2281	47505825,821	810,507	Yellow
5	22523	1398	29826887,652	533,587	Yellow
6	545	1254	29042065,063	514,109	Yellow
7	14068	1371	28012430,119	469,598	Yellow
8	22505	1170	20067881,908	331,108	Yellow
9	14967	976	19291122,805	329,885	Yellow
10	24221	769	19223242,914	348,411	Yellow
11	25661	1005	17021386,193	299,571	Yellow
12	16159	755	16104989,300	284,249	Yellow
13	5851	666	15961600,632	289,482	Yellow
14	24998	902	15862302,609	274,717	Yellow
15	22906	1038	15417655,342	275,801	Yellow
16	19894	877	13985011,630	227,515	Yellow
17	8326	519	12212543,545	218,540	Yellow
18	17100	650	12043258,345	212,823	Yellow
19	21756	657	11713411,487	199,097	Yellow
20	18259	677	11418587,429	199,302	Yellow
21	9453	442	10757209,738	194,210	Yellow
22	20675	565	10658914,761	190,439	Yellow
23	10613	801	10315637,653	202,944	Yellow
24	11661	419	9709931,999	172,401	Green
25	21281	615	9532653,353	174,276	Green
26	24853	398	9392758,095	166,607	Green
27	5729	450	9209548,481	161,384	Green



Facebook Network Graph: Distributions in Facebook

Degree distribution of node degree in the network

- Social Networks usually follow power-law distributions, such as $P(k) \sim k^{-\gamma}$, with k node degree and $\gamma \leq 3$.
- 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 (CCDF).

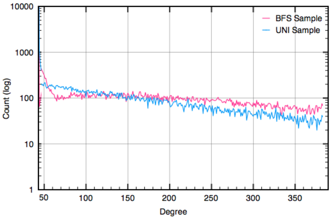


Figure: Tail of the power-law

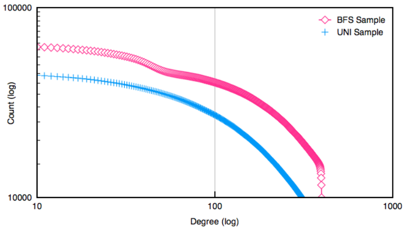


Figure: CCDF

Facebook Network Graph: Graph clustering

Clustering coefficient Is the measure representing how much nodes of a graph tend to “group” each other.

Results The mean value detected in FB lies in the interval [0.05, 0.2], the same w.r.t. other well-known real Social Networks.

Diameter of the network The mean diameter is smaller than 10, such as in the Milgram Small World theory.

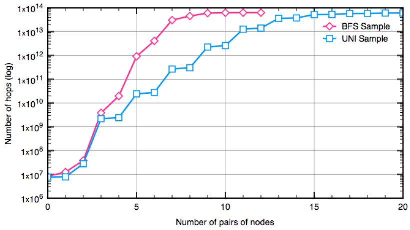


Figure: Diameter

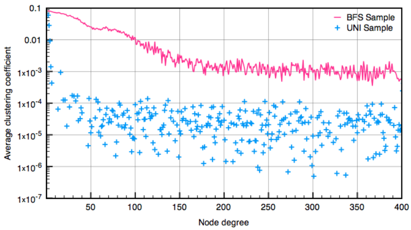


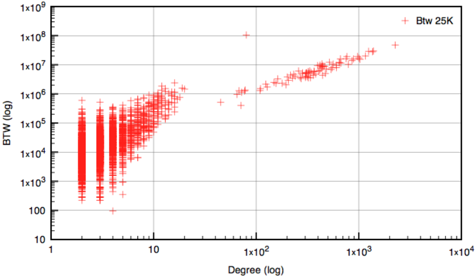
Figure: Clustering coefficient

Facebook Network Graph: Betweenness centrality

Betweenness centrality b_i of a node i is defined as $b_i = \sum_{j \neq k} d_{ij} \frac{n_{jk}(i)}{n_{jk}}$, i.e., the number of times the nodes lies in the shortest path connecting two other nodes.

Remarks It is well-known that the BC follows a power-law $p(g) \sim g^{-\eta}$ in scale-free networks.

Results We proved that it holds also for the Facebook network.



Facebook Community Structure

Community Structure A sub-structure of the overall graph, in which the density of relationships within the community is much greater than the density of connections among communities.

Model A common formulation of this problem is to find a partitioning $V = (V_1 \cup V_2 \cup \cdots \cup V_n)$ of disjoint subsets of vertices of the graph $G = (V, E)$ representing the network, in a meaningful manner.

Algorithms The most popular quantitative technique is the $Q - \textit{modularity}$ (or network modularity), proposed by Newman [4].

Q-modularity

$$Q = \sum_{s=1}^{m} \left[\frac{l_s^2}{E} - \frac{d_s^2}{2E} \right] \tag{1}$$

l_s: number of edges between vertices belonging to the s-th community; d_s: sum of the degrees of these vertices.

High values of Q [0,1] implies a evident community strucure.

[4] Finding and evaluating community structure in networks, Newman, 2004

Facebook Network Graph: Algorithms and Results

LPA (Label Propagation Algorithm) [5]

FNCA (Fast Network Community Algorithm) [6]

Algorithm	N. of Communities	Q	Time (s)
BFS (8.21 M vertices, 12.58 M edges)			
FNCA	50,156	0.6867	5.97e+004
LPA	48,750	0.6963	2.27e+004
Uniform (7.69 M vertices, 7.84 M edges)			
FNCA	40,700	0.9650	3.77e+004
LPA	48,022	0.9749	2.32e+004

Table: Results on Facebook Network Samples

[5] Near linear algorithm to detect community structures, Raghavan et al., 2007

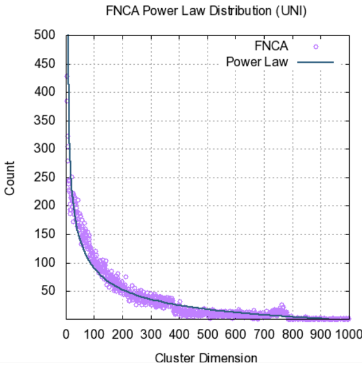
[6] Fast Complex Network Clustering Algorithm Using Agents, Jin et al., 2009

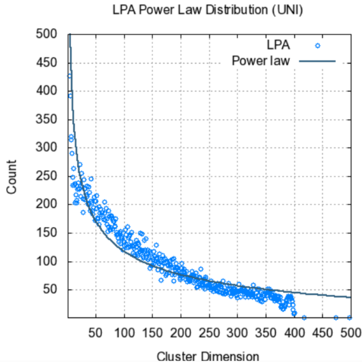
Facebook Community Structure: Uniform Sample

Uniform Sample The power law distribution is evident.

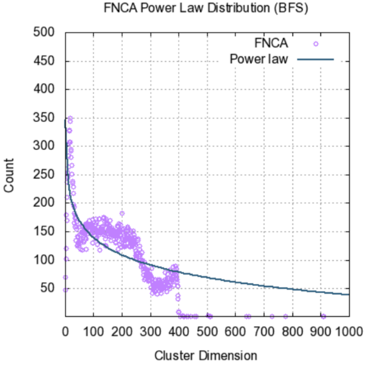
FNCA Algorithm $P(k)_{FNCA} \sim k^{-\gamma}$, with k node degree and $\gamma = 0.53$.

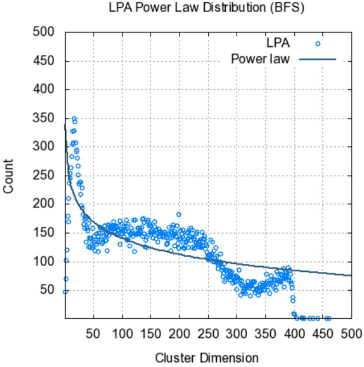
LPA Algorithm $P(k)_{LPA} \sim k^{-\gamma}$, $\gamma = 0.49$.





Facebook Community Structure: BFS Sample





- The differences in the behavior between the BFS and “Uniform” samples distributions reflect accordingly with the adopted sampling techniques.
- Gjoka et al. [7] and Kurant et al. [8], put into evidence the possible bias introduced by using the BFS algorithm, towards high degree nodes.

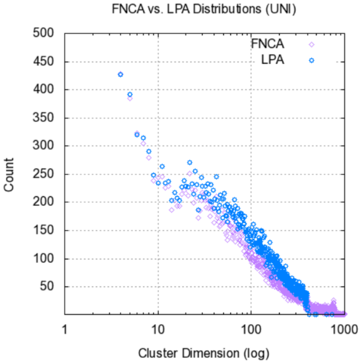
[7] Walking in facebook: A case study of unbiased sampling, Gjoka et al., 2010
[8] On the bias of BFS (Breadth First Search), Kurant et al., 2010

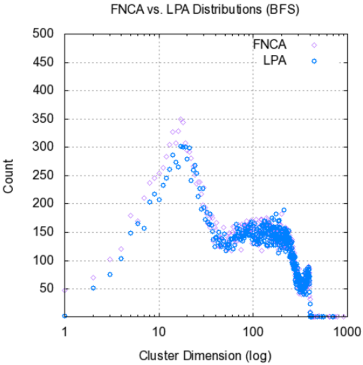
Facebook Community Structure: Overlapping Distributions

Observation These two distributions, regardless the sampling and community detecting adopted algorithms, appears to be strongly overlapping.

Q: We would qualitatively investigate the similarity among producted results, w.r.t. LPA and FNCA techniques.

A: We could compare obtained sets using similarity metrics, e.g., Jaccard and/or Cosine Similarity.





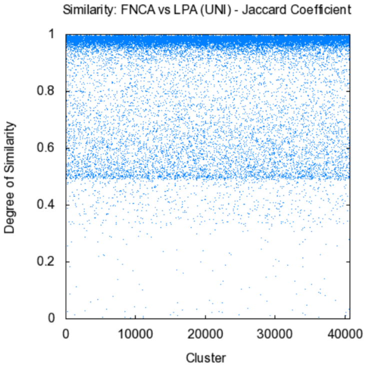
Facebook Community Structure: Similarity Measures

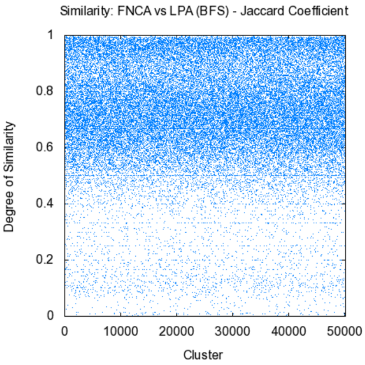
- Binary Jaccard Coefficient: $$\hat{J}(\mathbf{v}, \mathbf{w}) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$
  where M_{11} represents the total number of shared elements between vectors **v** and **w**, M_{01} represents the total number of elements belonging to **w** and not belonging to **v**, and, finally M_{10} the vice-versa.
  Intuitively, the result lies in $[0, 1]$.
- Cosine Similarity: $$cos(\Theta) = \frac{A \cdot B}{||A|| \, ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$
  where A_i and B_i represent the binary frequency vectors computed on the list members over i.

		Degree of Similarity FNCA vs. LPA			
Metric	Dataset	In Common	Mean	Median	Std. D.
$\hat{J}$	BFS	2.45%	73.28%	74.24%	18.76%
	Uniform	35.57%	91.53%	98.63%	15.98%

Table: Similarity degree of community structures

Facebook Community Structure: Similarity Results





Related Work I

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: International Conference 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

Related Work II

- Catanese S., De Meo P., Ferrara E., Fiumara G., Provetti A.
  Extraction and Analysis of Facebook Friendship Relations.
  In: Social Network Book 2011 (under review).

- Catanese S., Ferrara E., Fiumara G.
  Social network analysis of Facebook.
  In: Journal of Computational Science Special Issue on Social Computational Systems (under review).

- Ferrara E., Fiumara G., Baumgartner R.
  Web data extraction, applications and techniques: A survey.
  In: ACM Computing Surveys (under review).

- Ferrara E.
  Community Structure Discovery in Facebook
  In: Int. J. Social Network Mining (under review).

- Ferrara E., Jin D.
  A Large-Scale Community Structure Analysis in Facebook
  In: 11th International Conference on Knowledge Management and Knowledge Technologies (under review).