Extracting Collective Intelligence from Social Content

Yiannis Kompatsiaris
Multimedia Knowledge Lab
CERTH - Informatics and Telematics Institute
Thessaloniki, Greece
http://mklab.iti.gr

weknowit

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Contents

- Defining Collective Intelligence
- Collective Intelligence in WeKnowIt
- Two examples
  - Clustering in social content
  - Community detection in social content
- Conclusions - Issues

Defining Collective Intelligence

"Collective Intelligence is the ***INTELLIGENCE*** of a ***COLLECTIVE***, which arises from a number of ***SOURCES***"

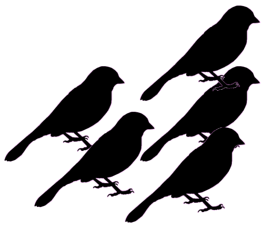
...an ***INTELLIGENCE*** that an individual cannot achieve by itself

What is collective?

Human Collectives

(e.g. communities, groups, organizations, families)

Groups of intelligent agents in computer environments



Animal Collectives

(e.g. ants, birds, bees)

Who is intelligent?

There are too many different definitions out there.

Defining intelligence is controversial and elusive activity.

Characteristics, capacities, functions that can be ascribed to intelligence

problem solving

decision making

applying knowledge

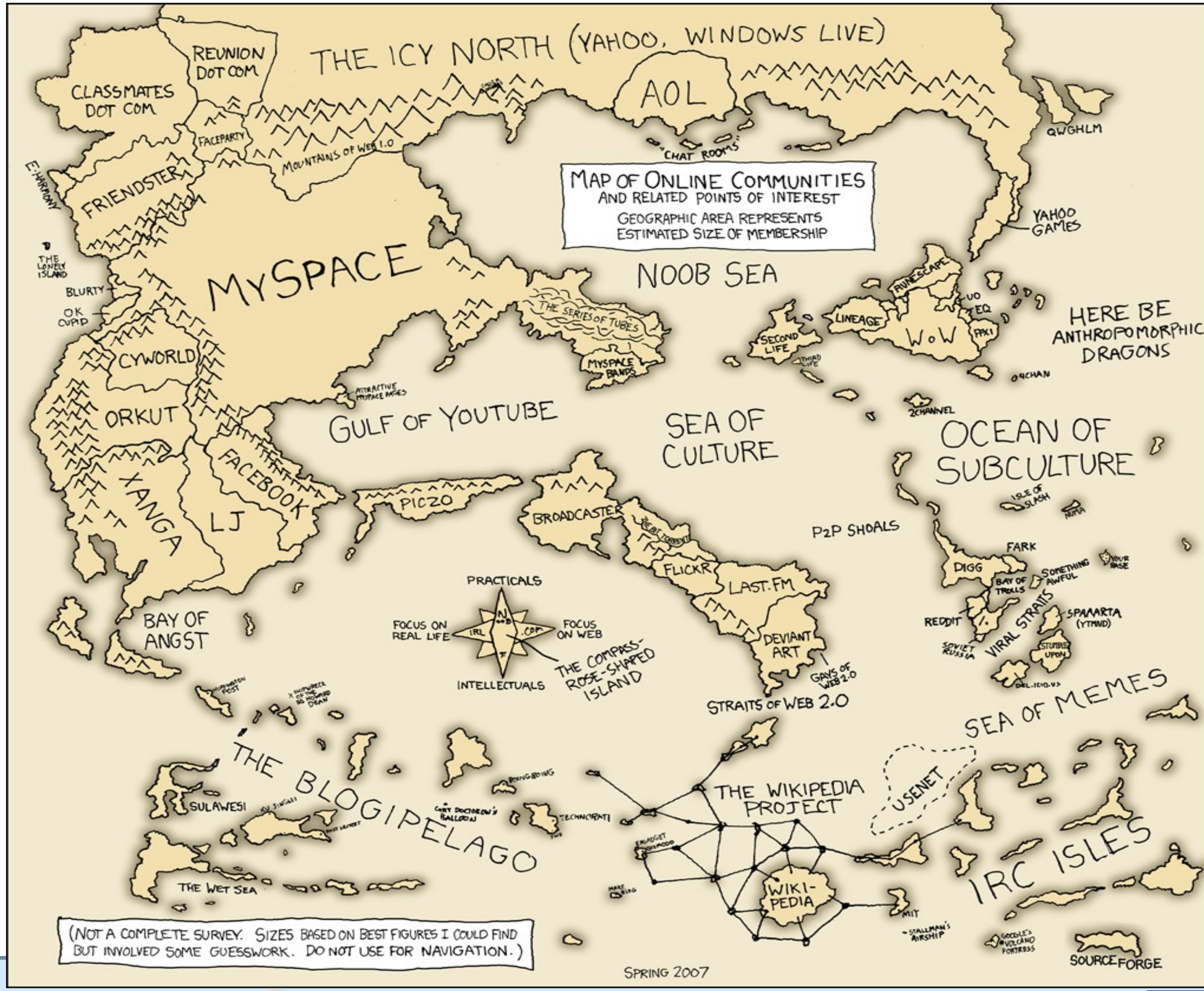
integration, synthesis

reasoning

information gathering,
sorting and categorization

evolution

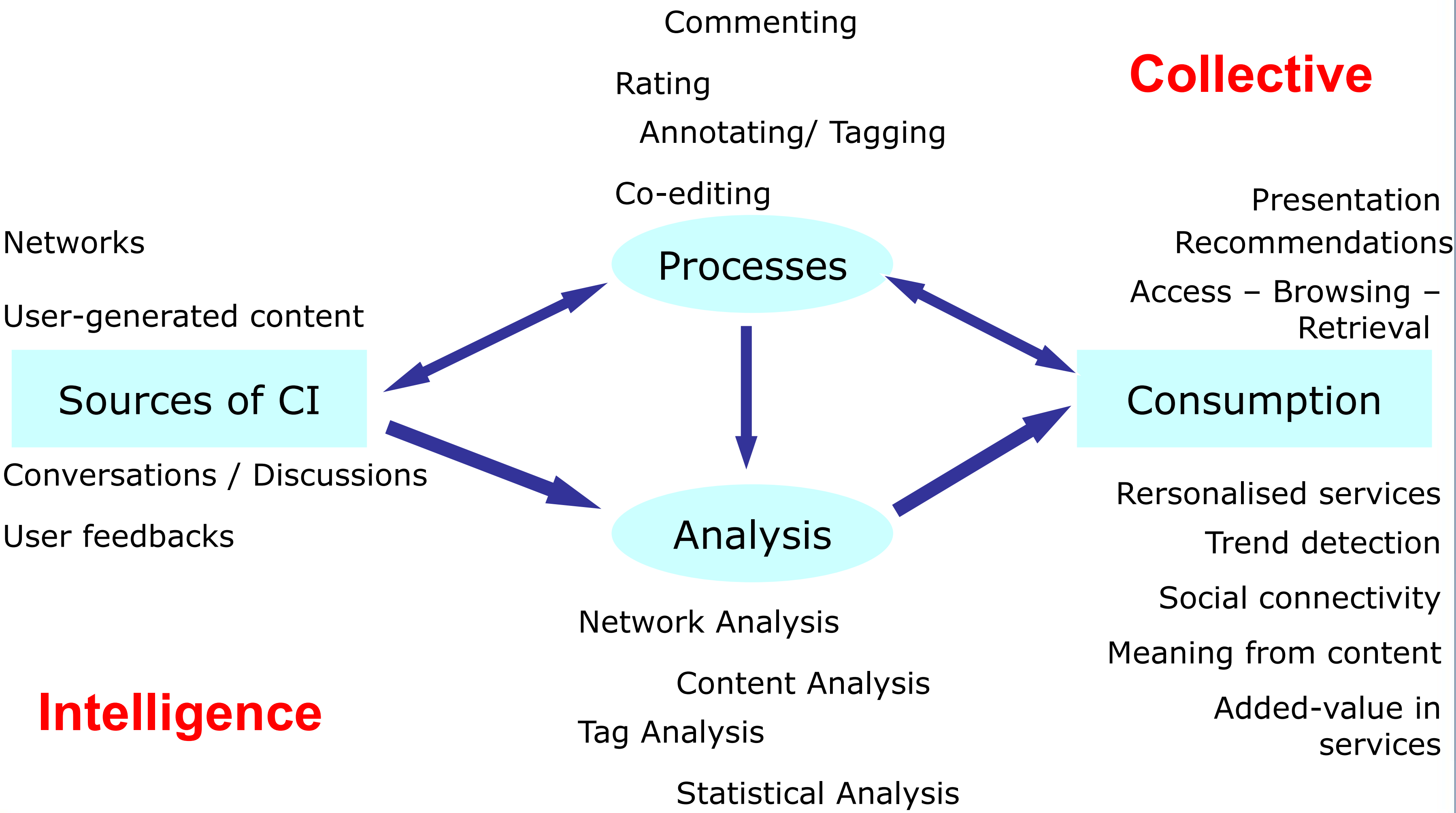
Why today?



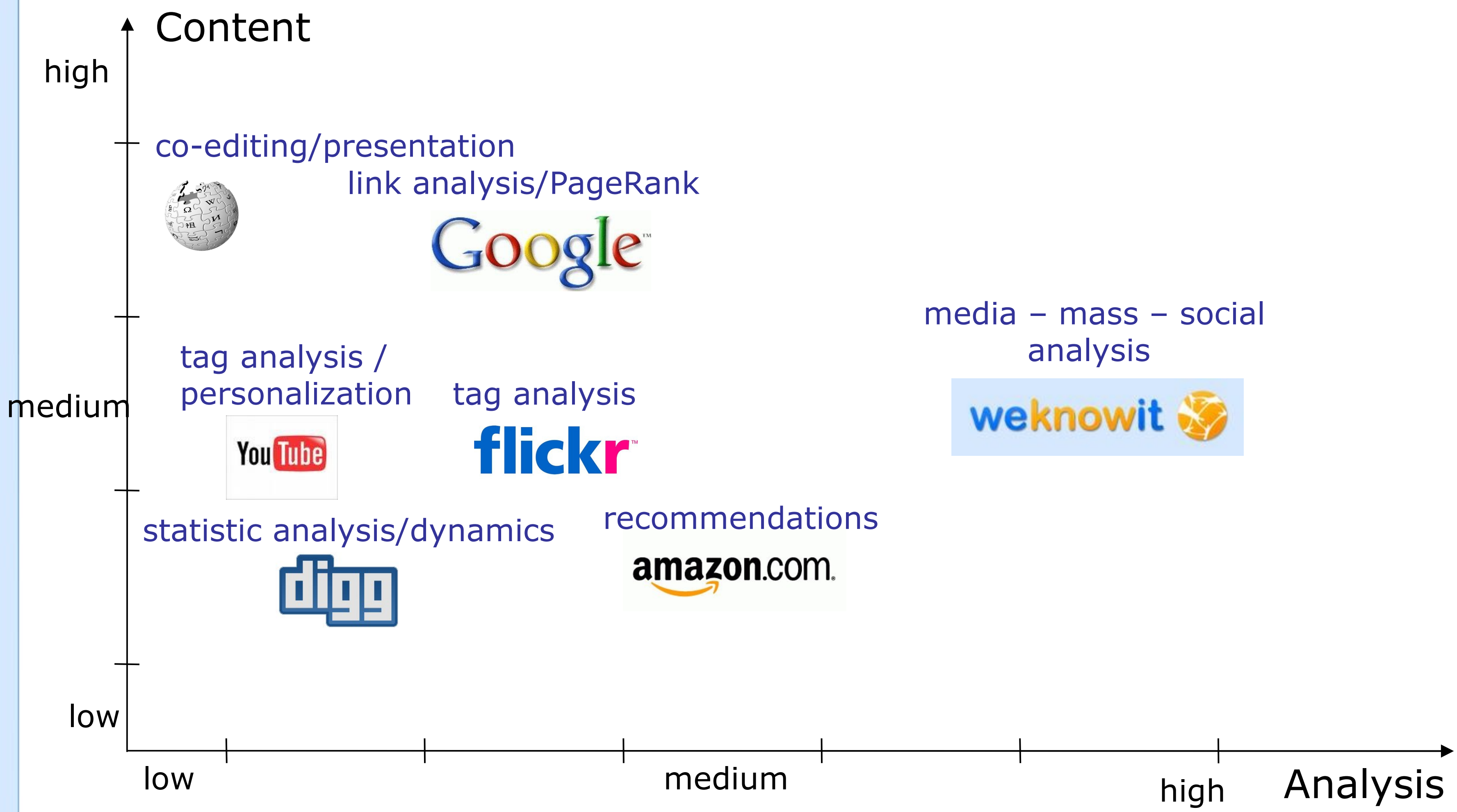
SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Collective Intelligence Overview

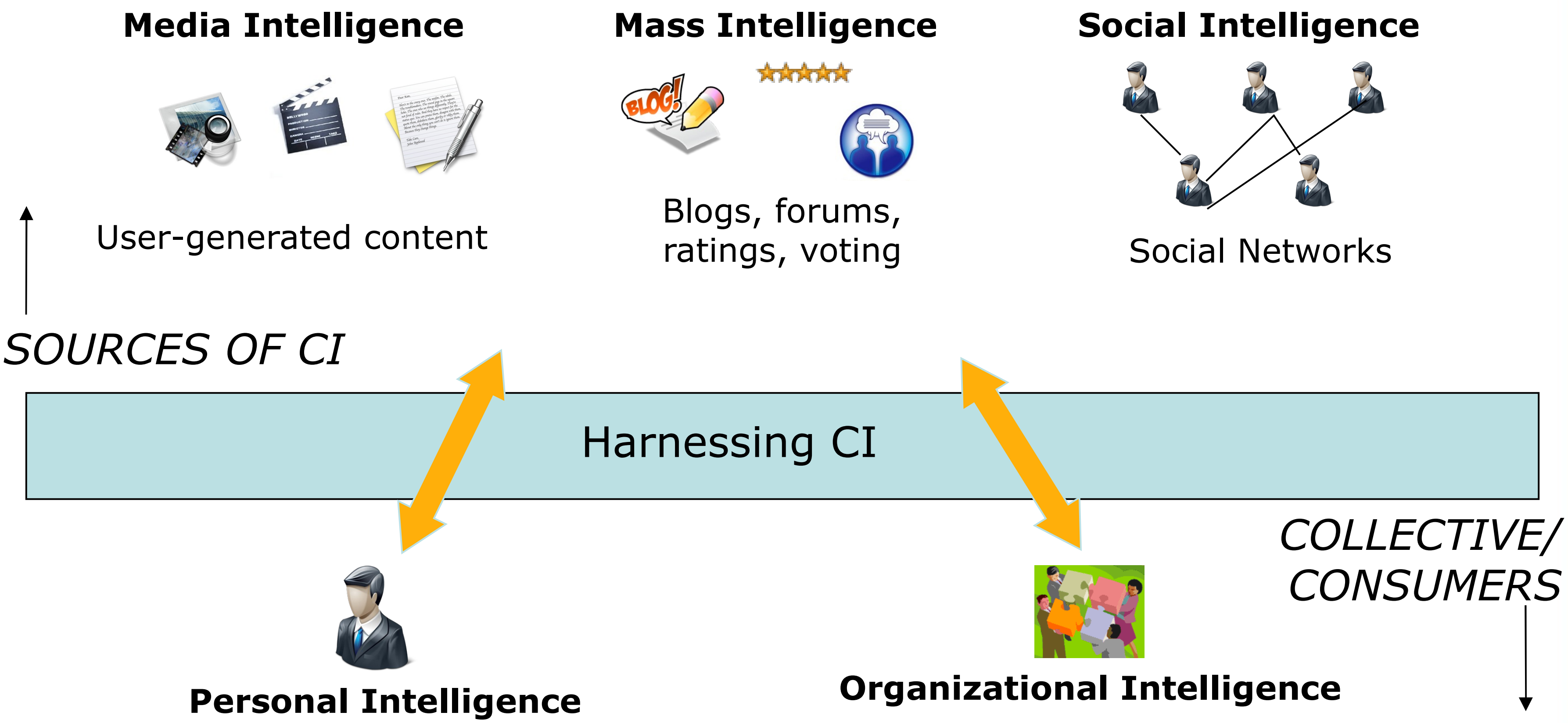


Analysis? What analysis?



WeKnowIt and CI

Decomposition of Collective Intelligence



weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

WeKnowIt and CI

Progress beyond State-of-the art Knowledge applications

- Personal Intelligence

(e.g. Amazon Recommendations)

- Media Intelligence

(e.g. TrecVid challenge)

- Mass Intelligence

(e.g. PageRank, Flickr, YouTube)

- Social Intelligence

(e.g. LinkedIn)

- Organizational Intelligence

(e.g. MS SharePoint)

WeKnowIt

Knowledge applications are based on combinations of these five layers of intelligence.

Harness the capabilities of truly Collective Intelligence!

Harnessing CI @ WeKnowIt

Personal Intelligence

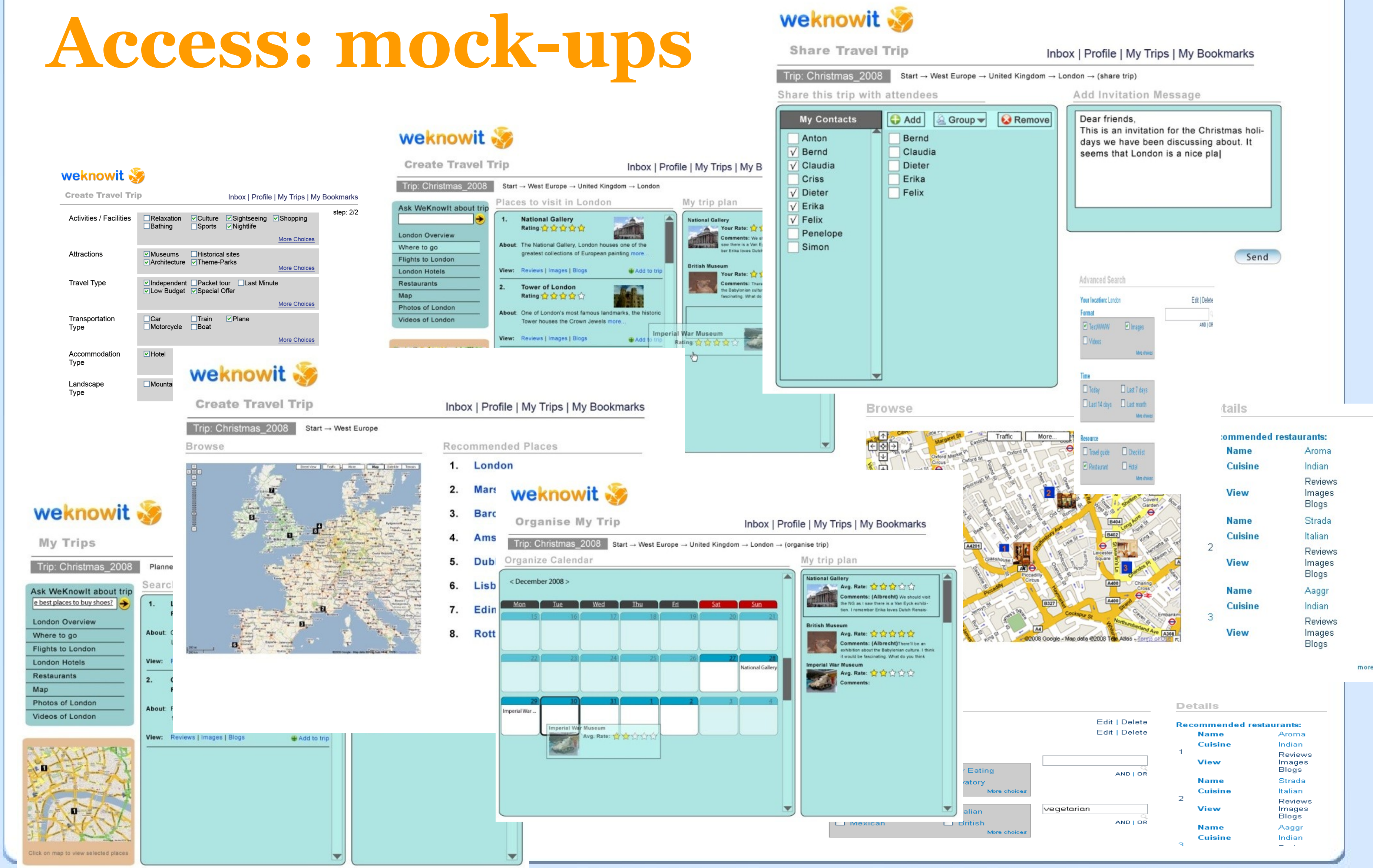
"enable personalized effective and efficient interaction with the applications"

The User is one part of the whole. One member of the COLLECTIVE

WeKnowIt aims at harnessing the individual intelligence and provide personalized services to the user

- Modeling and extraction of users preferences
- Efficient upload of user content
- Personalized access to content and generated intelligence

Access: mock-ups



SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Harnessing CI @ WeKnowIt

Media Intelligence

"Knowledge and information extraction from raw content in conjunction with contextual information, personal and social context"

Intelligent Content Analysis: fusing information from

diverse modalities (video/image, audio, text)

contextual information (location, time)

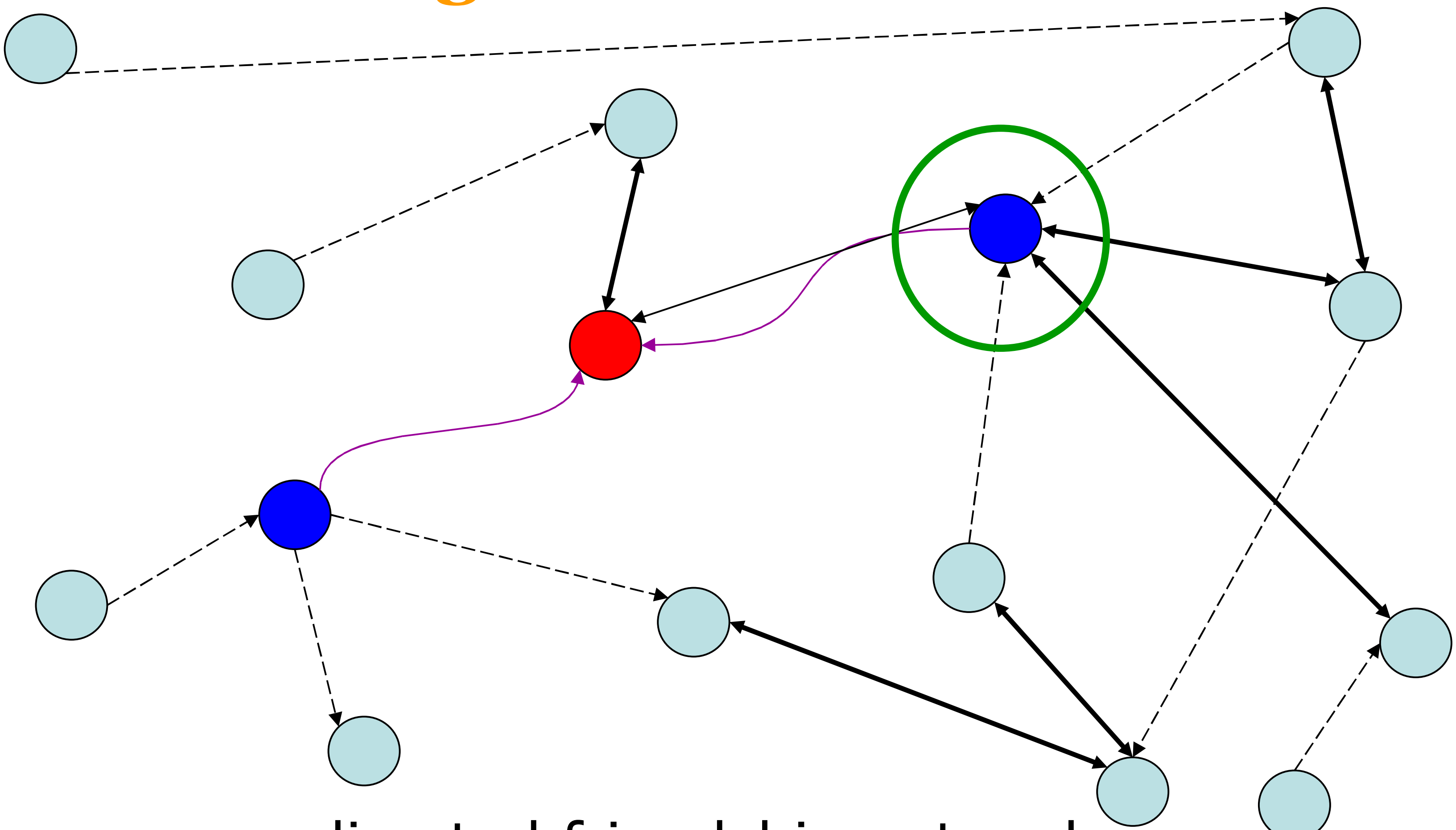
personal Context (user profile)

\+ **social Context** (friends, communities, tags, related items)

= fusion task, semantic analysis of content

A Source of Collective Intelligence

Which source to trust?
Social Intelligence



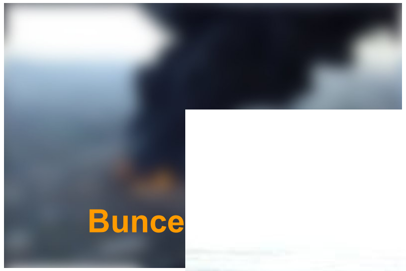
directed friendship network

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Social Network Analysis

- Individuals (actors) are not isolates regarding their actions. They always act within the possibilities and constraints given by their social environment
- Examples
  - Smoking in groups of high school kids
  - Fashion
  - Trading at the stock market
- Interactions are modelled as networks
- Methods from such fields as graph theory, mathematics, physics, sociology, social psychology are used to analyze these networks

Personal Intelligence

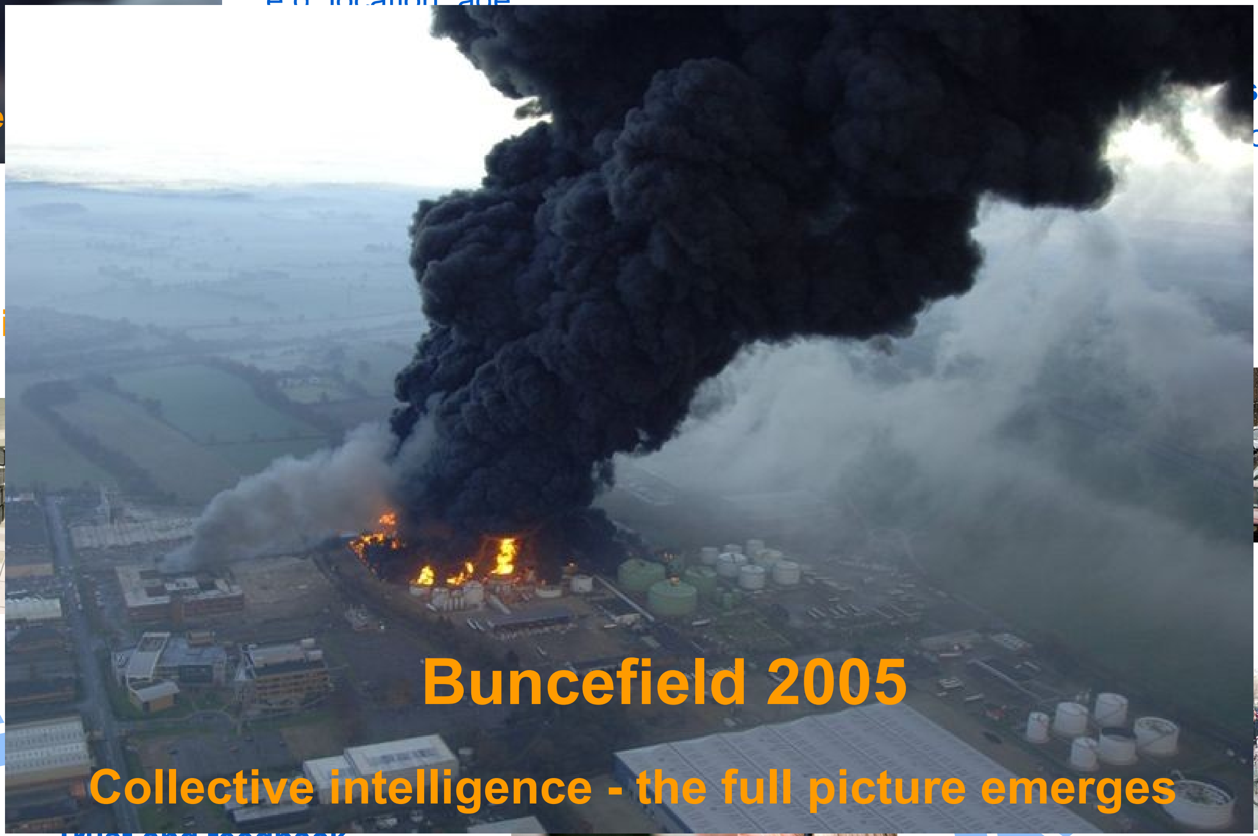


Profile of contributor
>> What to send where,
e.g. location, age

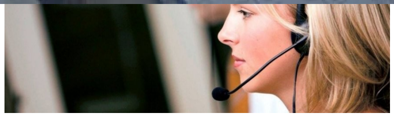
Media Intelligence

se
on

Organizati



Trust and feedback
>> Determine trustworthiness and hub-structures by SNA



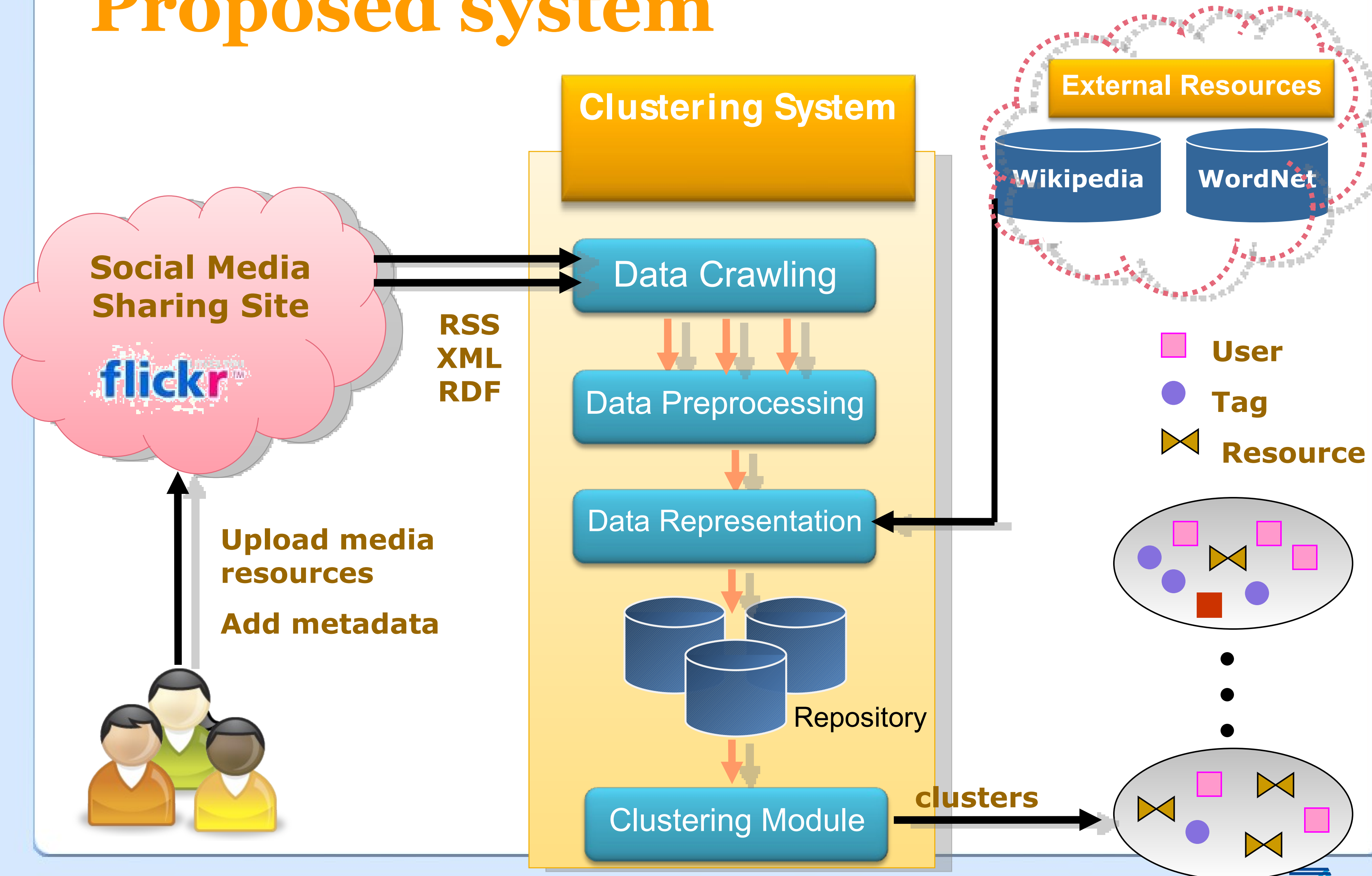
Social Tagging & Multimedia Content Clustering

- **Background**
  - High availability of multimedia content in social media sharing sites as source of CI
  - Plenty of user-generated metadata
  - Stable patterns in tagging systems over time
- **Motivation**
  - Poor IR (lack of structure of information, tag polysemy/ambiguity, chaotic environment)
  - Questionable tag validity
- **Problem Formulation**
  - Overcome of limitations and exploitation of (hidden) knowledge harvested in social media sharing sites through **clustering**.

Clustering Approaches

- Tag-Based
- Content-Based
- Co-clustering
  - Tags - resources
  - Time-based: users and tags

Proposed system



Tag-based Clustering (I)

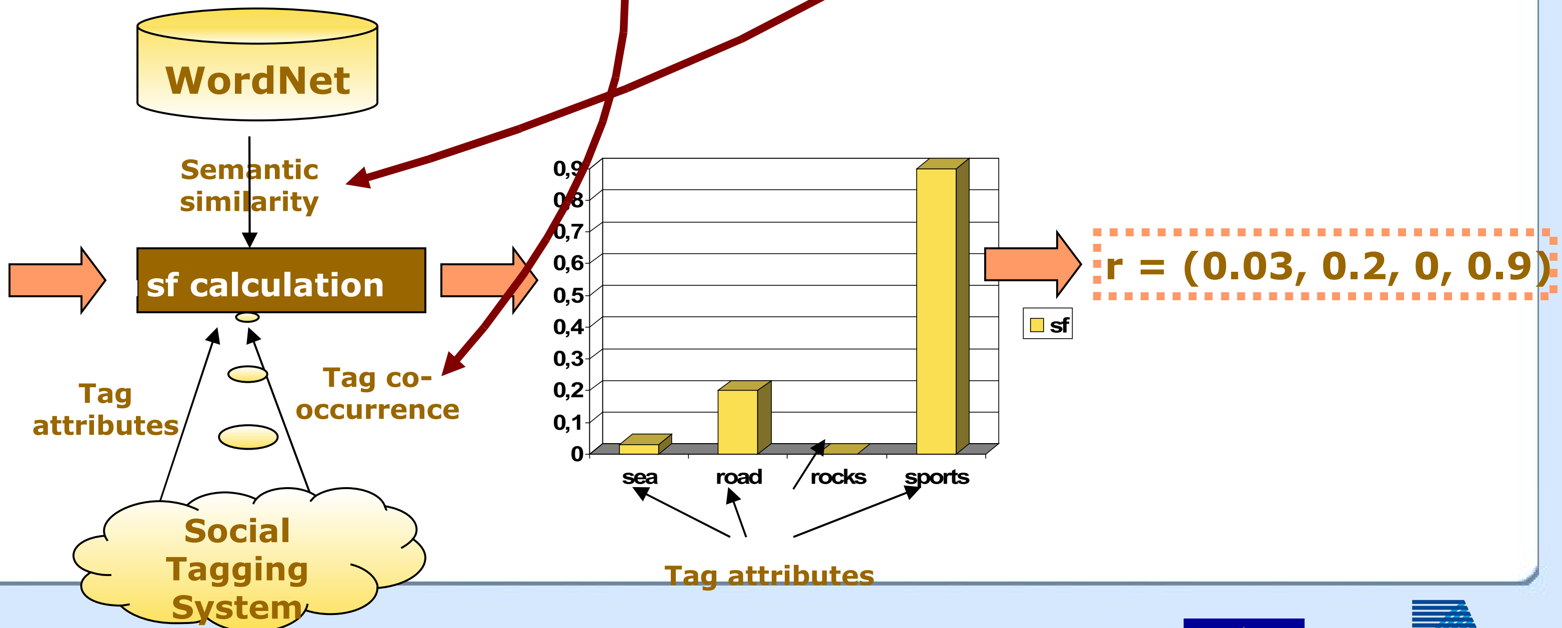
- **1. Vector data model**
- Assume **n** resources and **d** attribute-tags
  - **d:** a representative set of tags
- A resource representation in vector space (**sf**) is based on semantic similarity and tag co-occurrence between the resource's tags and the attribute-tags
- A resource $\mathbf{r_i}$ is represented by a **d**-dimensional vector $\mathbf{r_i} = (\mathbf{sf_1}, \mathbf{sf_2}, \ldots, \mathbf{sf_d})$
- All resources can be represented by an **n** x **d** matrix

Tag-based Clustering (II)

- **2. Clustering on n (resources, r) x d (attributes) matrix (K-means, Hierarchical, COBWEB)**

$$SS(t_x, t_y) = w * SoS(t_x, t_y) + (1 - w) * SeS(t_x, t_y)$$

Tennis, Roland Garros 2005



SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Tag-based Clustering - Experimental Results

- **Dataset:** 3000 images downloaded from Flickr
- Meaningful subdomains of **roadside**:

buildings, roof, street, road

(a)

cars, vehicles, race

(b)

people, street, festival

(c)

- Different clusters for the **ambiguous tag** *wave, rock:*

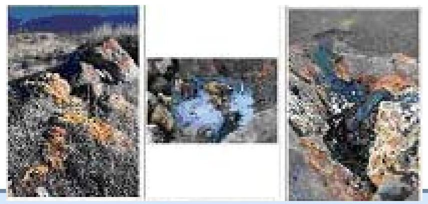
wave, sea, ocean

(a)

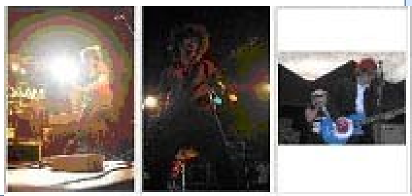
wave, person, hand

(b)

rocks, stone, rockyside



rock, music, band

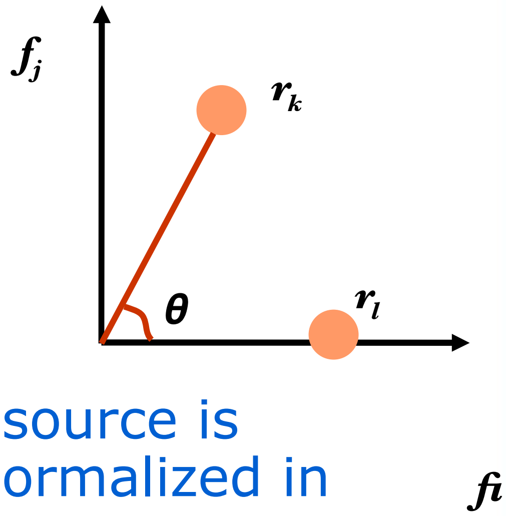


weknowit

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Tag & Content-based Clustering

- **Method:** After performing tag-based clustering, low-level features of resources are used for cluster refinement (outlier detection)
- **Vector data model**
- For each resource the following visual descriptors are extracted:
  - Scalable Color, ***SC***
  - Color Structure, ***CS***
  - Color Layout, ***CL***
  - Edge Histogram, ***EH***
  - Homogenous Texture, ***HT***

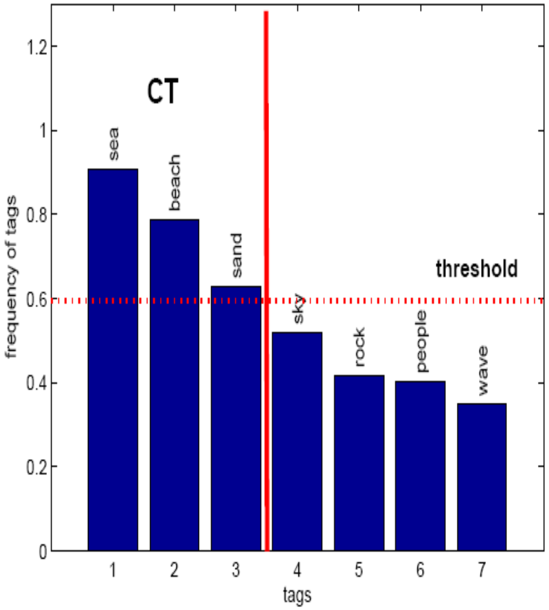


- A single image feature vector per each resource is produced, encompassing all descriptors normalized in [0,1]
- Feature extraction and distances between image feature vectors are according to MPEG-7 XM.

Evaluation Method

- **Definition:** Cluster Topic, CT, are the tags that have frequency in cluster's resources annotation over a threshold *T*.

- **Evaluation Metrics**
- Precision $$Pr(C_j) = \frac{|C_j \cap RR(C_j)|}{|C_j|}$$
- Recall $$R(C_j) = \frac{|RR(C_j) \cap C_j|}{|RR(C_j)|}$$
- F-Measure $$F(C_j) = \frac{2 * Pr(C_j) * R(C_j)}{Pr(C_j) + R(C_j)}$$

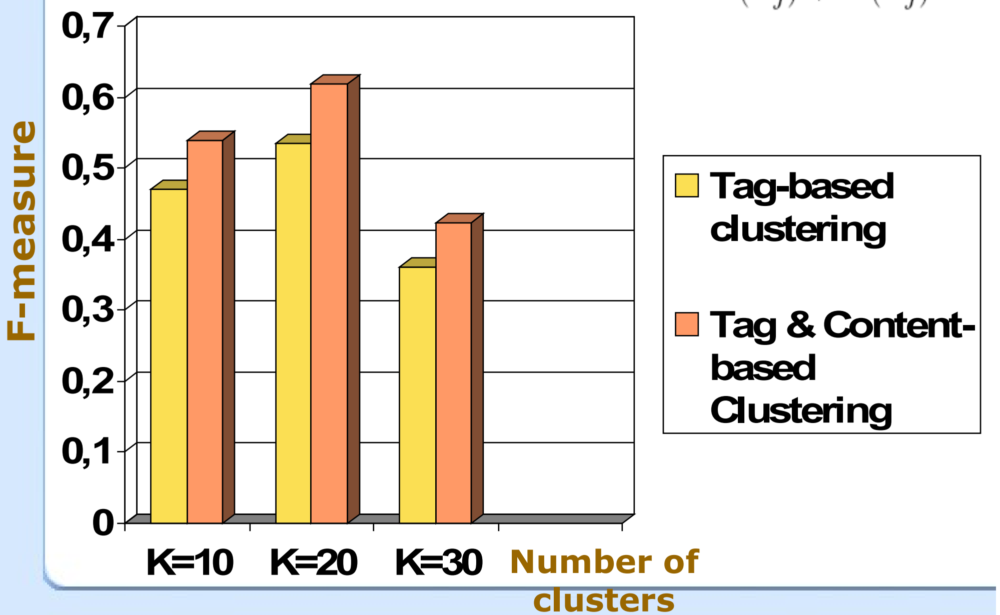


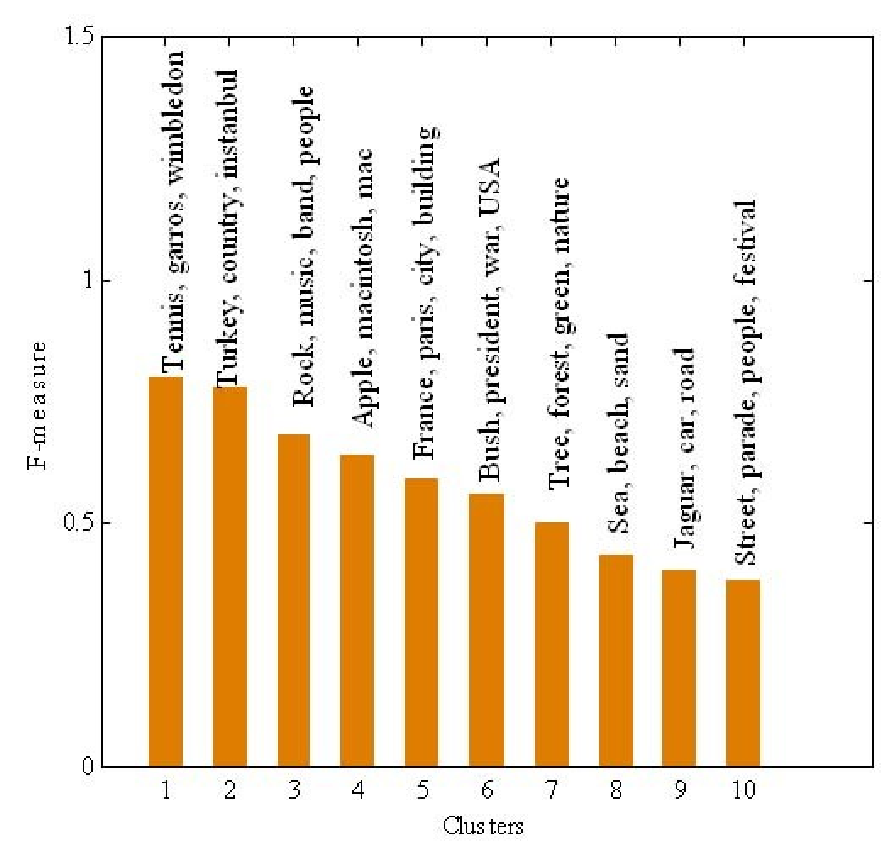
Tag & Content-based Clustering – Experimental Results

Dataset: 10000 images (with their tags) downloaded from Flickr

Evaluation: Manual annotation and use of F-Measure.

$$F(C_j) = \frac{2 * Pr(C_j) * R(C_j)}{Pr(C_j) + R(C_j)}$$

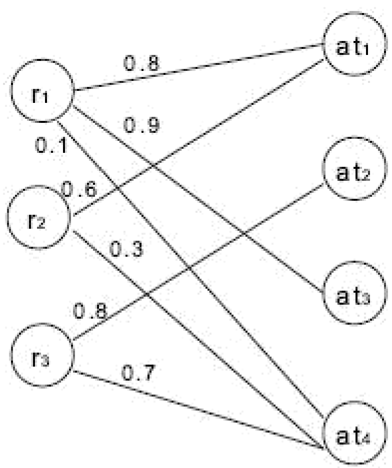




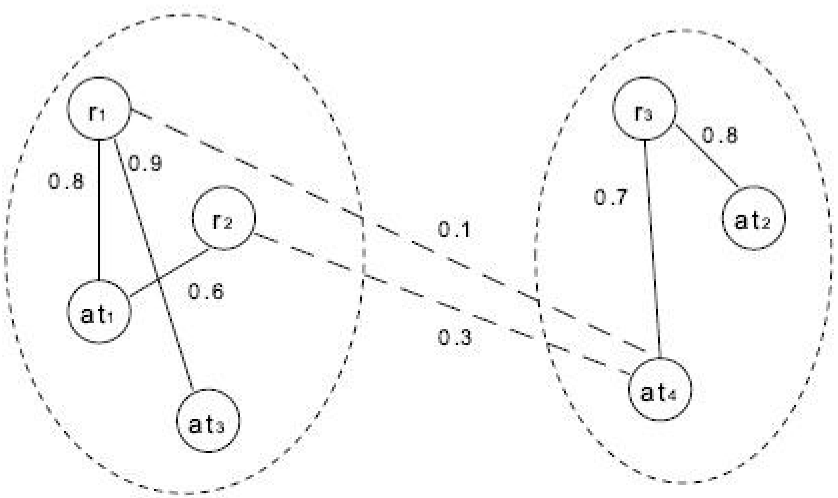
SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Co-clustering

- **Graph data model**
- A graph structure **G** = {**V$_1$**, **V$_2$**; **E**} is used for the representation of the dataset, where **V$_1$** and **V$_2$** can be sets of resources, users, tags or time intervals and **E** denotes the relations between the nodes of **V$_1$** and **V$_2$** .



Graph representation



Graph-partitioning problem

Co-clustering Tags & Resources

Problem: Find *k* clusters of both resources and tags, such that:

$$\sum_{x=1}^{k} \sum_{r_i, a_j \in C_x} Similarity(r_i, a_j), \forall\, r_i R, a_j AS$$

is maximized ■

R: Resources Set

AS: Tag-attributes Set

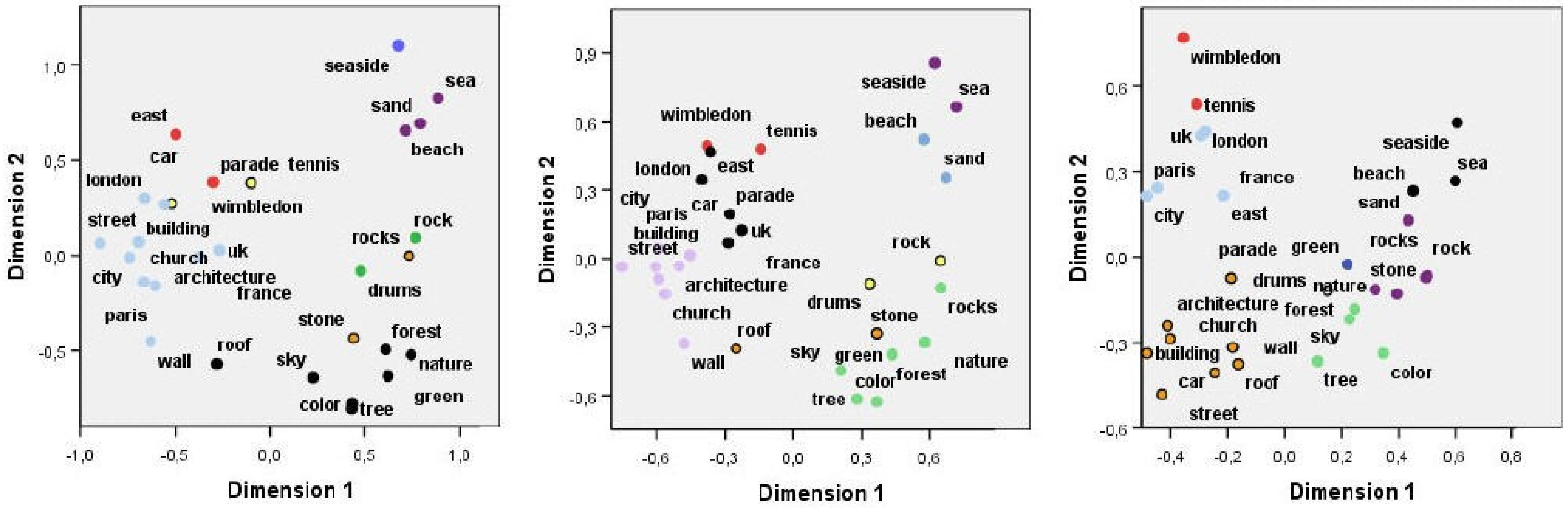
Algorithm 1 The CO-CLUSTERING algorithm.

Input: The set R of n resources, the set T of l tags and two integers k and w where $w \in [0..1]$

Output: A set $C = \{C_1, \ldots, C_k\}$ of k subsets consisting of elements from both R and T, such that the sum of inter-clusters similarities defined by (6) is minimized.

```
/*Preprocessing*/
T* = Preprocess(T)
AS = ExtractAttributes(T*)
/*capturing similarities*/
SoS = CalculateSocialSimilarity(R,AS)
SeS = CalculateSemanticSimilarity(R,AS)
SS = w * SoS + (1 − w) * SeS
RA = Similarity(SS)
/*Co-clustering process*/
(D_r, D_at) = ComputeDegreeTables(RA)
NRA = D_r^{-1/2} RA D_at^{-1/2}
(L_r, R_at) = SVD(NRA)
SV = CreateIntegratedTable(D_r, D_at, L_r, R_at)
C = k − means(SV, k)
```

Co-clustering Tags & Resources - Experimental Results (I)



(a) $w = 0.2$ (b) $w = 0.5$ (c) $w = 0.8$

Attributes Assignment to k=8 clusters,

W : weighting factor of semantic similarity against similarity derived from tag co-occurrence

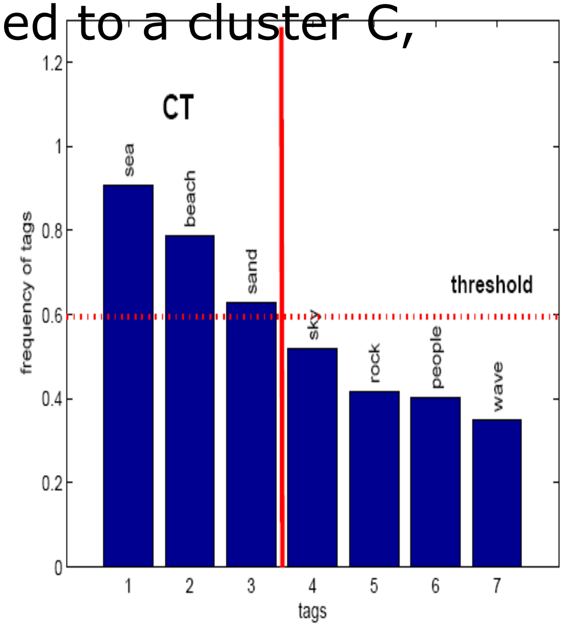
weknowit

SEVENTH FRAMEWORK PROGRAMME

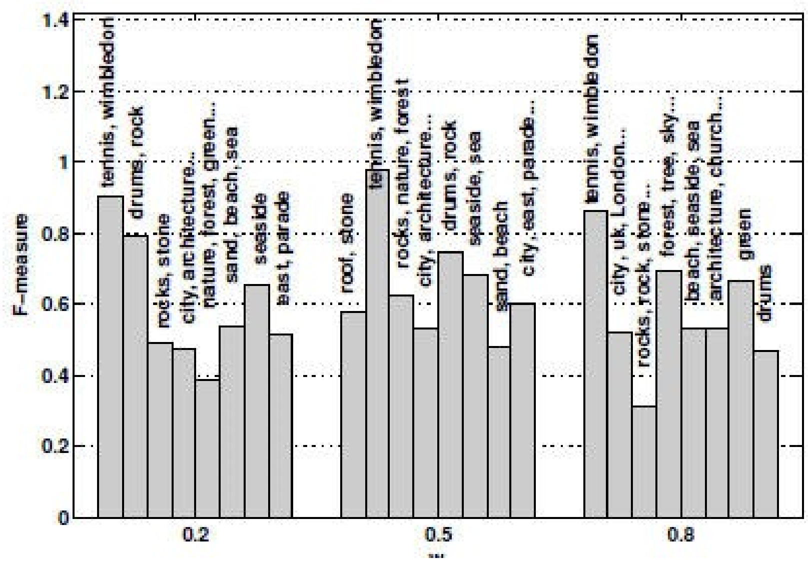
co-funded by the European Union

Co-clustering Tags & Resources - Evaluation Method

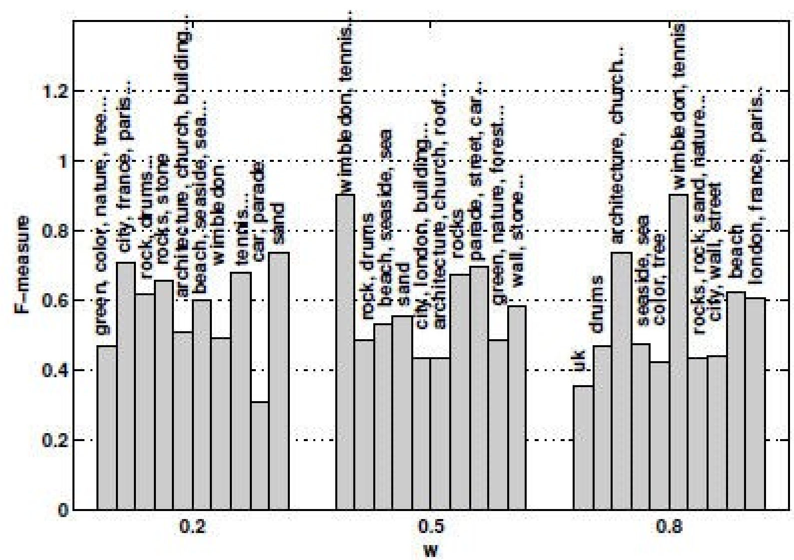
- **Definition:** Cluster Topic, CT, are the tags that have frequency in cluster's resources annotation over a threshold *τ*.
- A resource is considered correctly assigned to a cluster C, if it contains **all** the tags of the CT of C.
- **Evaluation Metrics**
- Precision $$Pr(C_j) = \frac{|C_j \cap RR(C_j)|}{|C_j|}$$
- Recall $$R(C_j) = \frac{|RR(C_j) \cap C_j|}{|RR(C_j)|}$$
- F-Measure $$F(C_j) = \frac{2 * Pr(C_j) * R(C_j)}{Pr(C_j) + R(C_j)}$$



Co-clustering Tags & Resources - Experimental Results (II)



k = 8

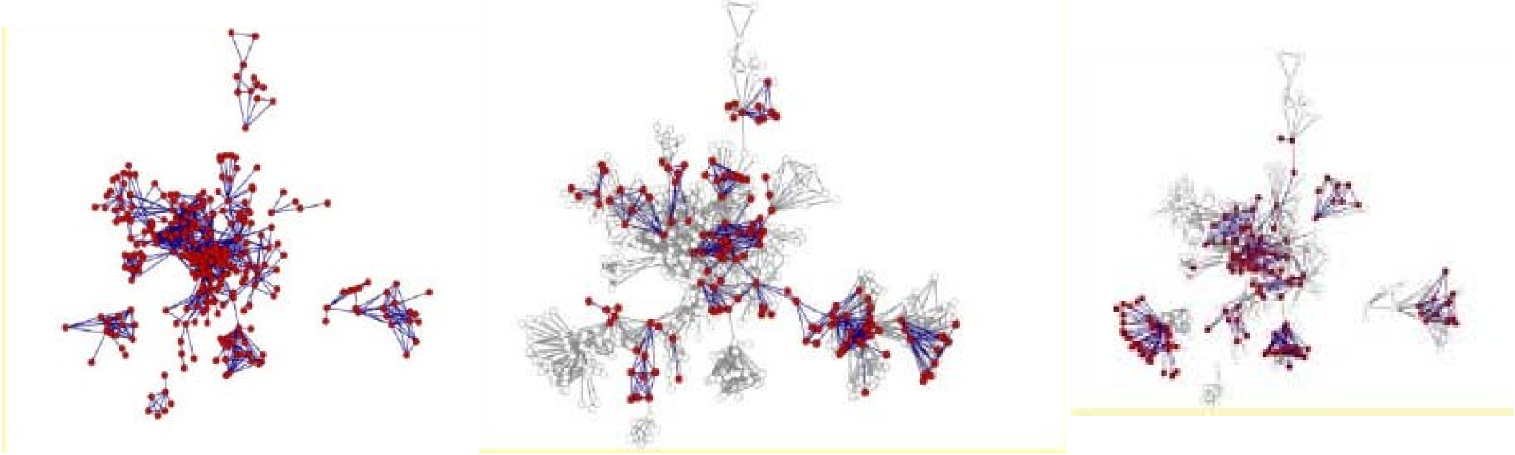


k = 10

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Users-Tags Co-clustering over time

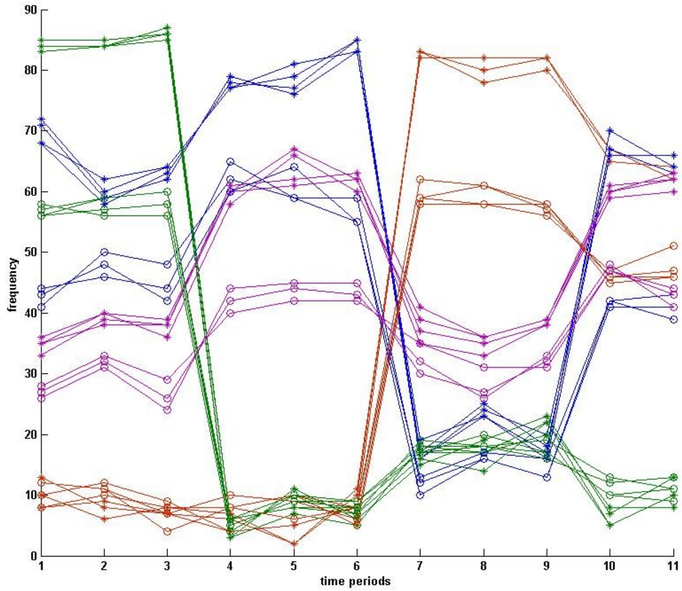
- **Problem:**
- Compute similarities over time between users and tags
- Find Dominating topics per time slot



Sample Co-Clustering Results

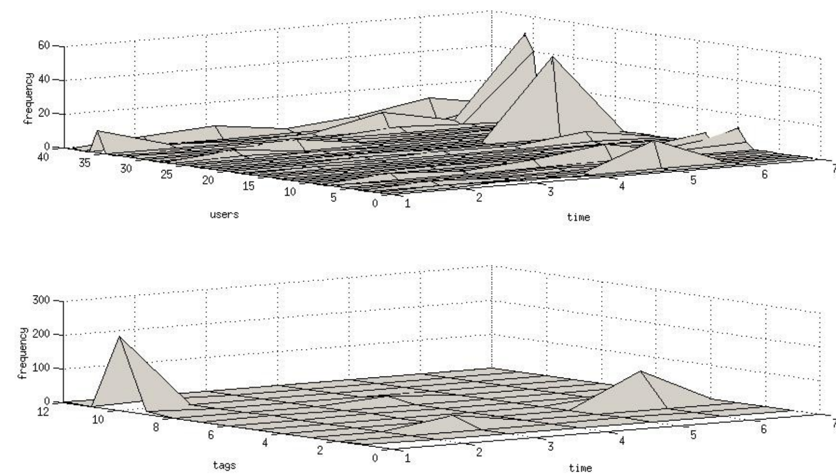
Synthetic data

- 15 users over a period of 11 time slots
- 14 tags over a period of 11 time slots



Real data

- Period: August 2007-August 2008
- Topics: earthquake, wedding, ancient Greece, Olympic games
- 1218 users, 4713 tags, 210 days



Conclusions

- Tag co-occurrence, semantic similarity of tags and content-based similarity of resources are useful indicators of IR in a social tagging system.
- Tag ambiguity, lack of structure and tag spamming can be sufficiently tackled.

- **Use Cases**
- Inducing ontology from Flickr tags (crawling, clustering, relationship extraction)
- Domain Ontology enrichment
- Social assisted analysis
- User profiles
- Recommendations
- Trend detection

Future work

- Improvement of Clustering Methods
- Testing of more Clustering Methods, Metrics, etc.
- Application of proposed use-cases
- Extension to geo-data analysis and clustering for social maps enrichment

Harnessing CI @ WeKnowIt

Mass Intelligence

"is recognition and understanding of facts and trends by exploitation of massive user contributions"

Sources of Collective Intelligence

Blogs (comments)

Forums (threads/discussions)

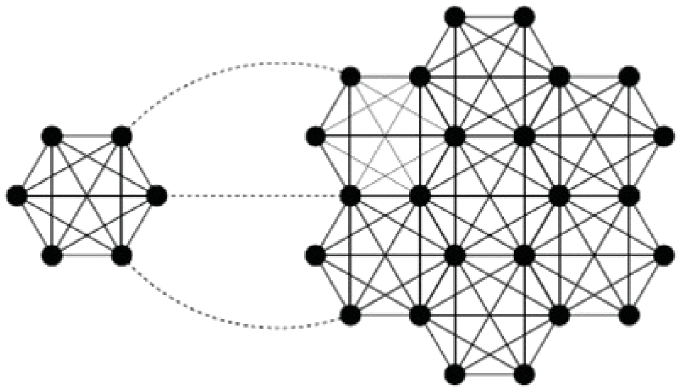
Ratings

Questions & Answers

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

Community Detection in Complex Networks

- Community Detection: The Problem
- Global vs. Local Community Detection
- Bridge Bounding
- Future Work



weknowit

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

Problem Statement

- No common definition of community.
- Some definitions:

  A community is a group of vertices with:

  - more edges among them than $\sum_{v \in C} w_{uv} \geq \sum_{v \in V-C} w_{uv}$ *for all* $u \in C$.

  between them and the rest of the graph,

  - high *modularity*, $Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$
  - high *conductance*. $\phi(S) = \frac{\sum_{i \in S, j \notin S} A_{ij}}{\min\{A(S), A(\overline{S})\}}$

- In any case, the output of a community detection process on a graph is a set of vertex sets.

Global vs. Local

- **Global:** Process the whole graph to derive a partition into communities
  + Abundant research
  + Good results (community quality, algorithm efficiency)
  - Not practical for huge graphs or for real-time applications
- **Local:** Incremental process of the graph and output communities (streaming)
  - Relatively little research
  - Great potential for demanding applications

Bridge Bounding

Algorithm

- Start a community with a seed node
- Add neighbouring nodes as long as they are connected by edges that are not inter-community ("bridges").
- Stop when it is not possible to add any more nodes.

Algorithm 1 LocalCommunityDetection

Require: Seed node $s \in G = (V, E)$
Require: Community mapping $g_C : V \rightarrow \mathbf{P}$
Require: Bridge function $b : E \rightarrow [0.0, 1.0]$

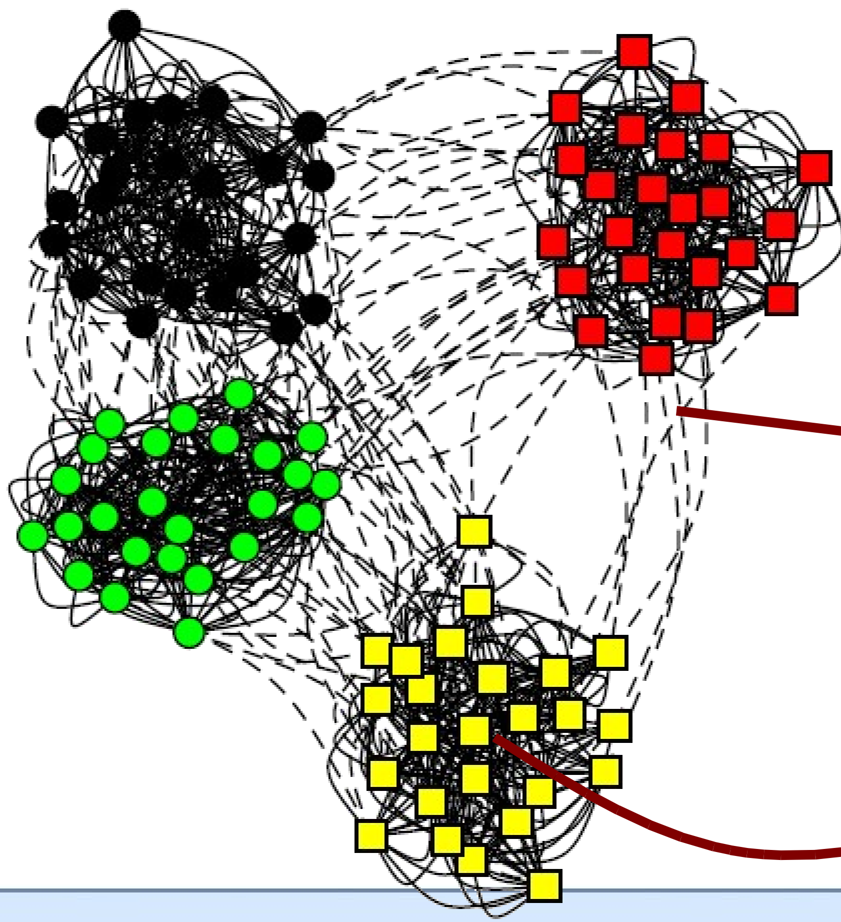
$C_s = \oslash$
Frontier set $F = \{s\}$
while $|F| > 0$ **do** $\{F$ is non-empty$\}$
  $c \leftarrow F.\mathrm{pop}()$
  $C_s \leftarrow C_s \bigcup \{c\}$
  $C_U \leftarrow C_U \setminus \{c\}$
  for all $n \in N(c)$ such that $e_{cn} = (c, n) \in E$ **do**
    if $g_C(n) = C_U$ and $b(e_{cn}) \leq B_L$ **then**
      $F.\mathrm{push}(n)$
    end if
  end for
end while
$\mathbf{P} \leftarrow \mathbf{P} \bigcup C_s$

Basic success factor:

Edge Bridge-ness: The property of an edge to lie between two communities.

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

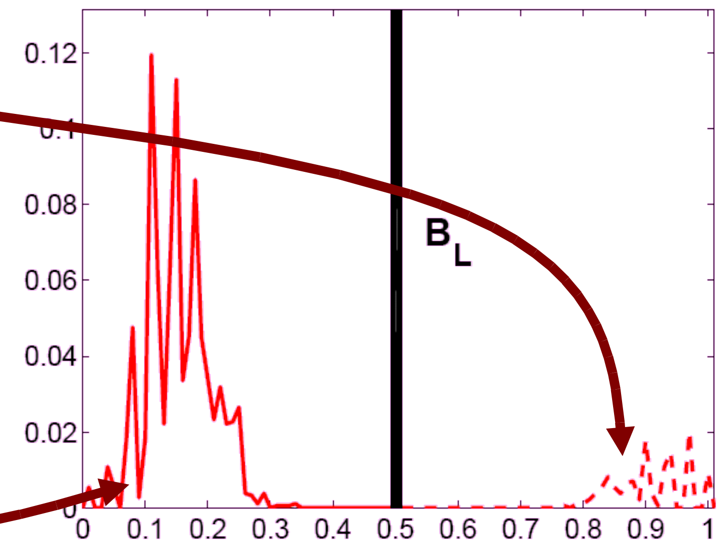
Bridge Bounding – Toy Example



Local bridging of an edge

$$b_L(e_{st}) = 1 - C_{st}^{(3)} = 1 - \frac{|N(s) \bigcap N(t)|}{min[(d(s)-1),(d(t)-1)]}$$

s, t: endpoints of edge
N(s), N(t): neighbourhoods of s, t
d(s), d(t): degrees of s, t



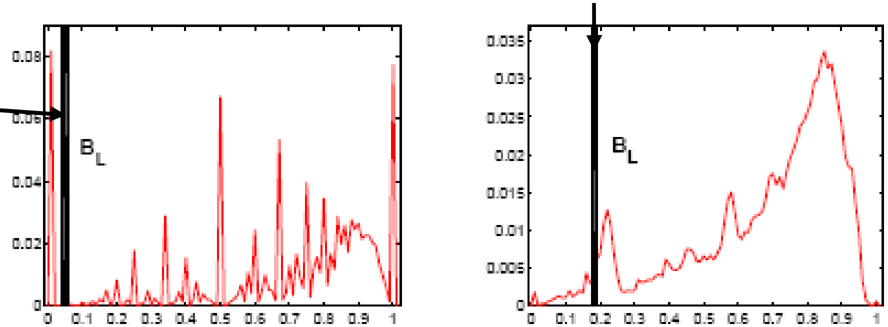
Bridge Bounding - Problems

- Local bridging not suitable for scale-free networks
- Solution (partial) 2nd order local bridging.

$$b'_L(e_{st}) = \alpha \cdot b_L(e_{st}) + (1-\alpha)\frac{1}{|N(e_{st})|}\sum_{e \in N(e_{st})} b_L(e)$$

B_L = 0.17 leaves just 1% of edges as non-bridges.

B_L as low as 0.05 leaves 8% of edges as non-bridges.

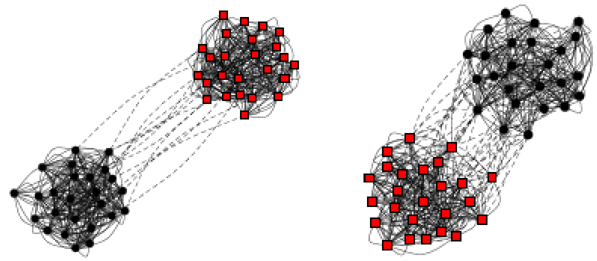


(a) b_L distribution (b) b'_L distribution, $\alpha = 0.7$

Experiments on Synthetic Community Networks

- Synthetic networks according to method of Newman and Girvan.

$$S_{PAR} = \{N, K, z_{tot}, p_{out}, s_{var}\}$$



(a) $p_{out} = 0.01$ (b) $p_{out} = 0.08$

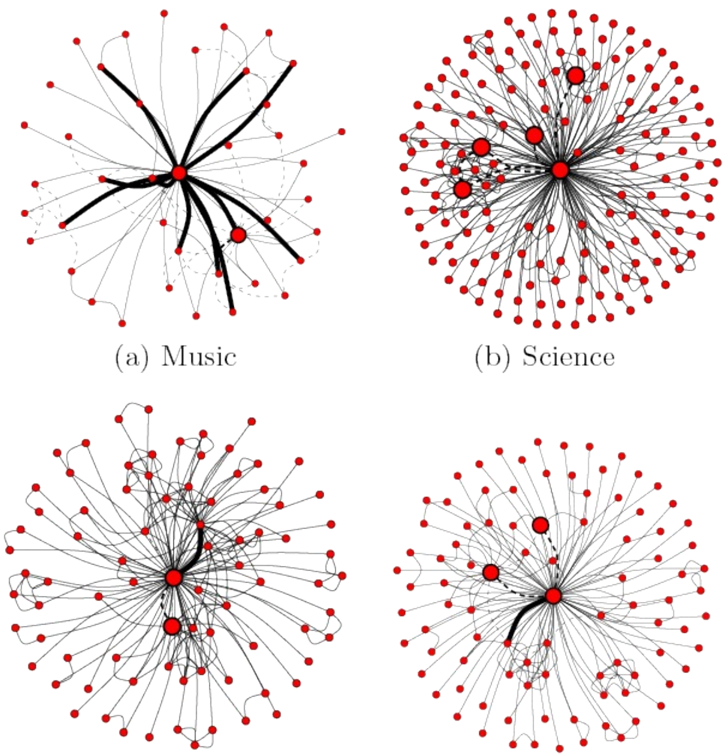
Change conspicuity of underlying communities.

	F_C			NMI		
p_{out}	BB	BB'	GN	BB	BB'	GN
0.01	100	100	100	1.0	1.0	1.0
0.05	100	100	100	1.0	1.0	1.0
0.1	100	100	50	1.0	1.0	0.86
0.15	100	99	50	1.0	.98	0.86
0.20	99	74	50	0.98	0.84	0.86
0.25	24	24	0	0.54	0.56	0.02

Change relative sizes of underlying communities.

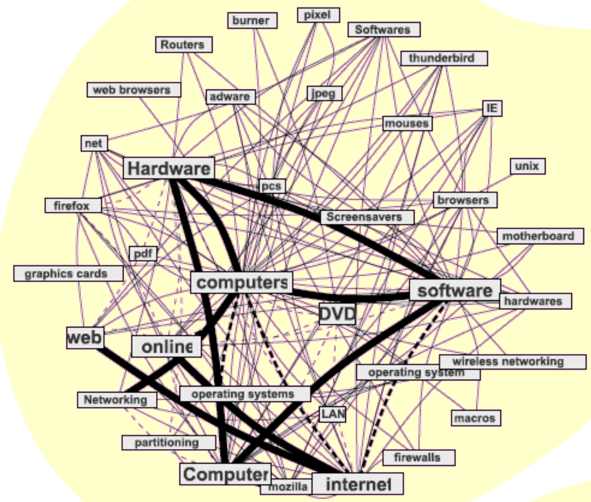
	F_C			NMI		
s_{var}	BB	BB'	GN	BB	BB'	GN
1.1	100	100	100	1.0	1.0	1.0
1.5	100	100	100	1.0	1.0	1.0
1.6	99.5	100	100	0.99	1.0	1.0
1.7	88	98	100	0.82	0.96	1.0
1.8	85.5	97	100	0.79	0.95	1.0
1.9	58.5	87	90	0.68	0.82	0.88
2.0	12.5	80	82	0.45	0.73	0.81
2.5	0	62	75	0.45	0.63	0.72

LYCOS iQ Tag Network

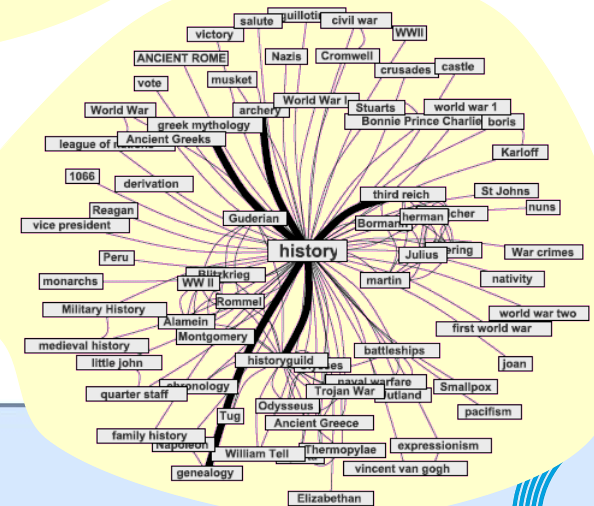


(a) Music (b) Science

(c) Film (d) Animals



Computers:
A densely interconnected community



History:
A star-shaped community

SEVENTH FRAMEWORK PROGRAMME

co-funded by the European Union

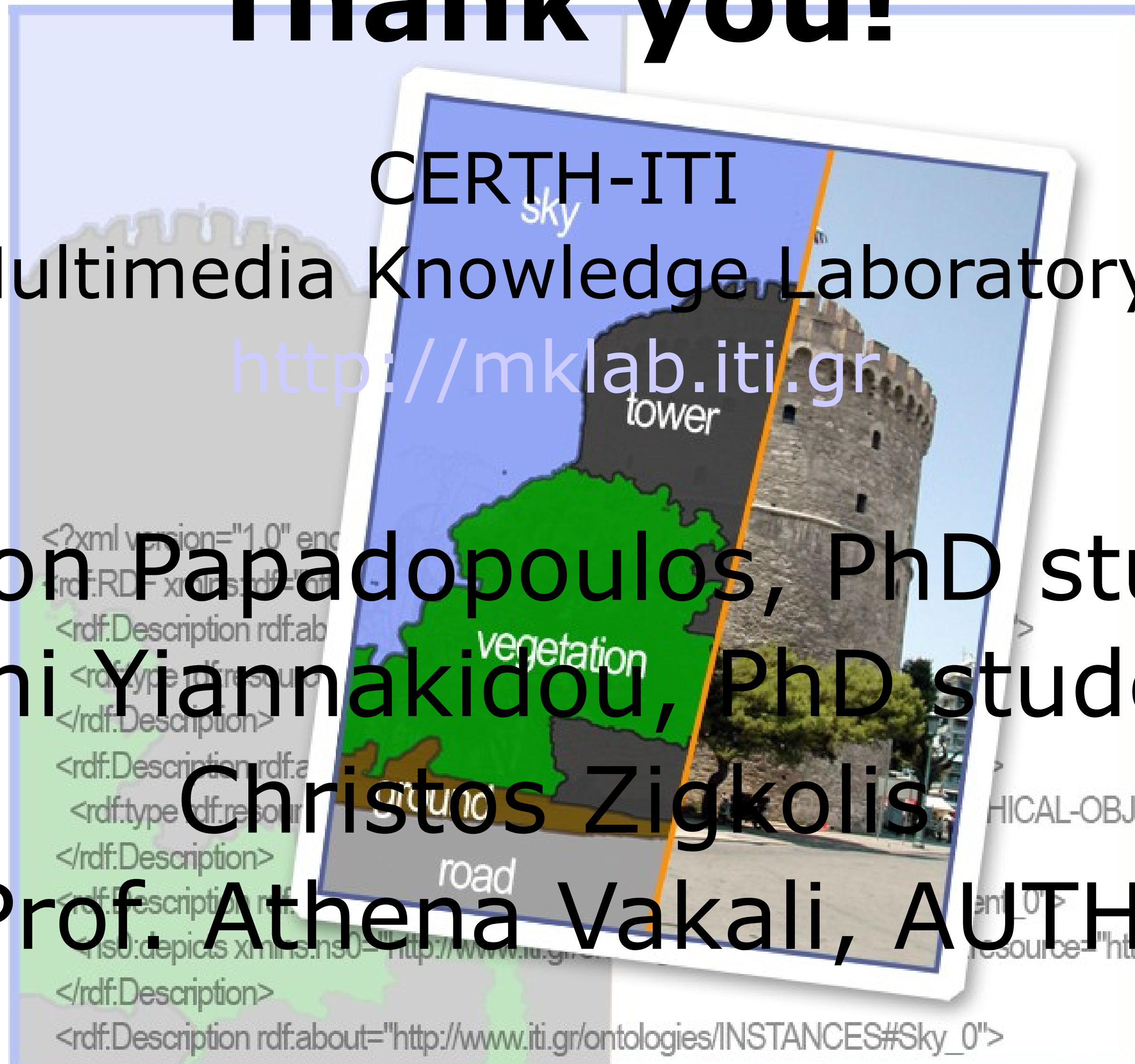
Future Work

- Remove ad-hoc parts of the algorithm:
  - Selection of B_L threshold.
  - Heuristics for artificially stopping community building process (e.g. co-occurrence frequency)
- Compare with other methods.
- Evaluate on real networks.
- Other applications

CI Issues

- Trust, security, privacy, wrong data
- Applications and commercialization
- Integration with services - organizations
- Efficiency of semantics and analysis
- Real integration
  - not just sum of different analysis
  - formal framework and approach
- User interaction
- Performance, scalability
  - speed, storage, power

Thank you!



CERTH-ITI
Multimedia Knowledge Laboratory
http://mklab.iti.gr

Symeon Papadopoulos, PhD student
Eirini Yiannakidou, PhD student
Christos Zigkolis
Prof. Athena Vakali, AUTH

weknowit

SEVENTH FRAMEWORK PROGRAMME
co-funded by the European Union

WeKnowIt Consortium

CERTH – ITI	⇨ Multimedia, Personalization, Management
UoKob	⇨ Collaborative Data Analysis, Knowledge Management
Lycos	⇨ Web 2.0 Platform, Data Provision, Mass Feedback
Telefonica	⇨ Personalization, Data Mining, Exploitation
USFD	⇨ Human-Computer Interaction, Text Analysis
EM-KA	⇨ Recommendation Systems, Social Networks
VOD	⇨ Mobile Service Provision
SMIND	⇨ Software Architecture & Integration, Exploitation
SCC	⇨ Emergency Response