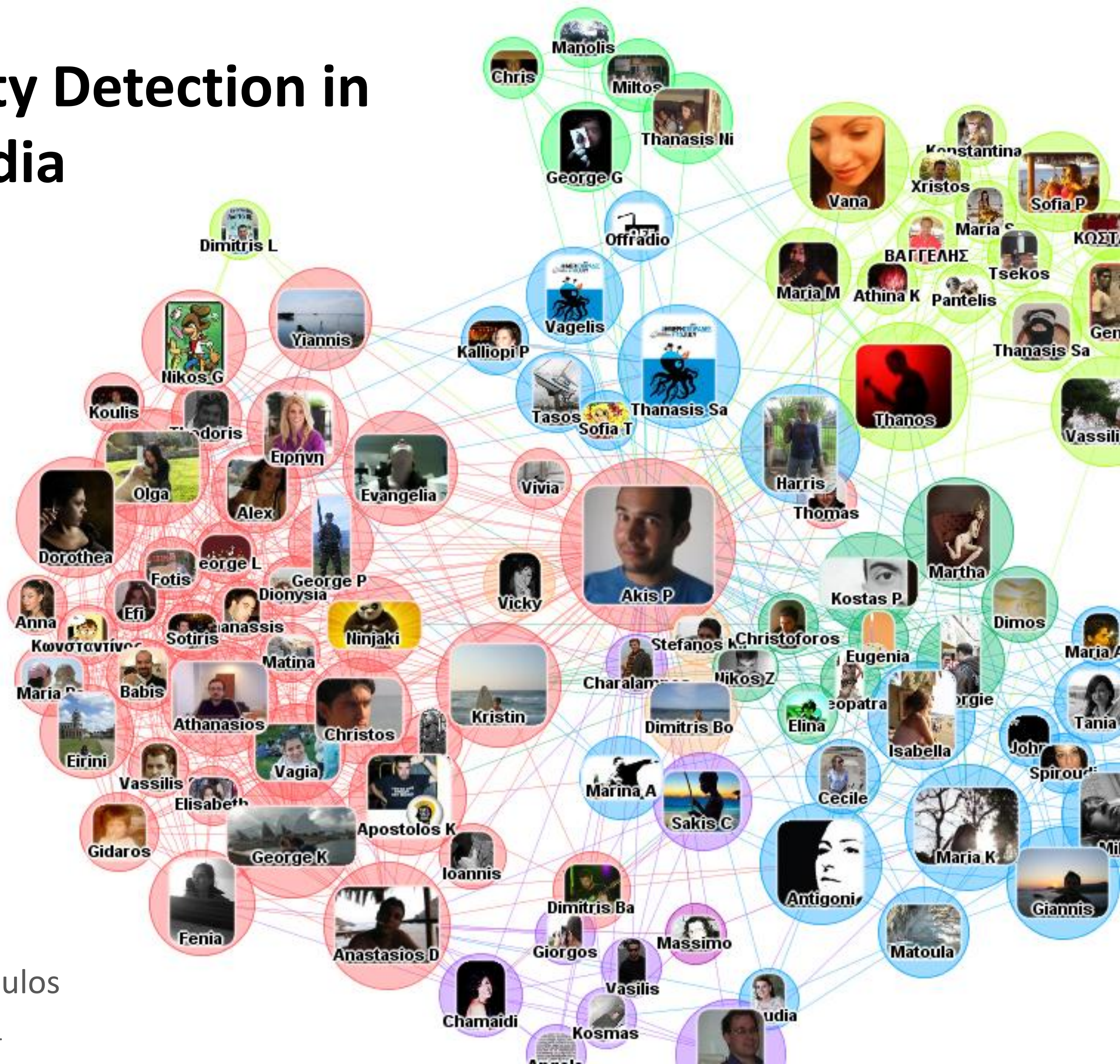
Community Detection in Social Media



Symeon Papadopoulos

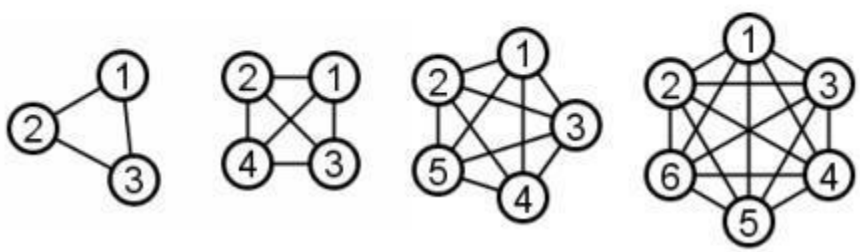
CERTH-ITI, 22 June 2011

graph

nodes

$$G = (V, E)$$

edges



graph basics

node degree

$\deg(v_i) = k_i$ = number of neighbors

In directed graphs, we differentiate between in- and out-degree.

adjacency matrix

A_{ij} = link between nodes *i* and *j*

0 → no link
1 → link
α → link with weight equal to α

graph construction from Web data *1*

Webpage www.x.com

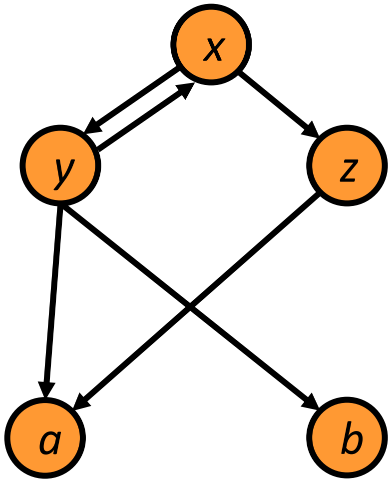
href="www.y.com"
href = "www.z.com"

Webpage www.y.com

href="www.x.com"
href = "www.a.com"
href = "www.b.com"

Webpage www.z.com

href="www.a.com"



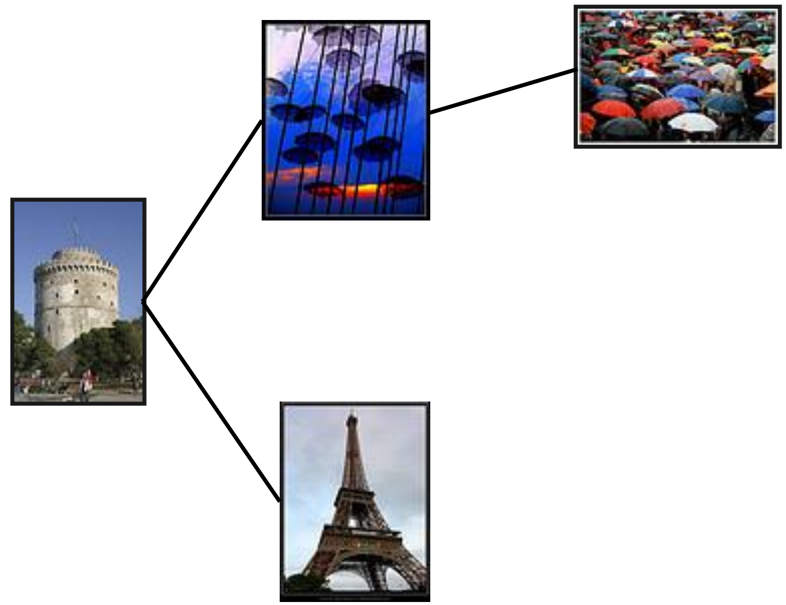
graph construction from Web data *2*

thessaloniki, umbrella

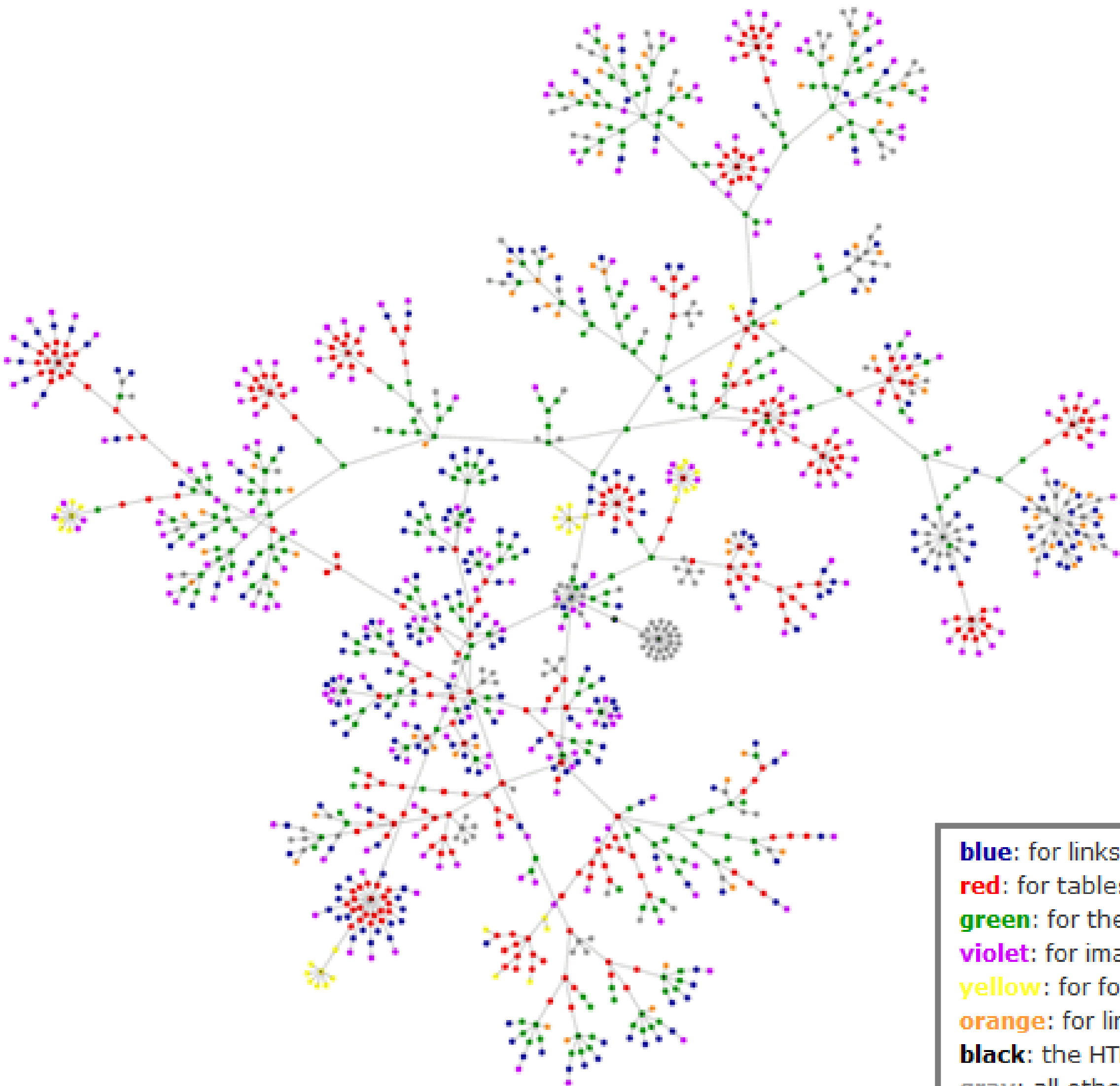
thessaloniki, tower

umbrella, crowd

eiffel, tower



web pages as graphs *1*

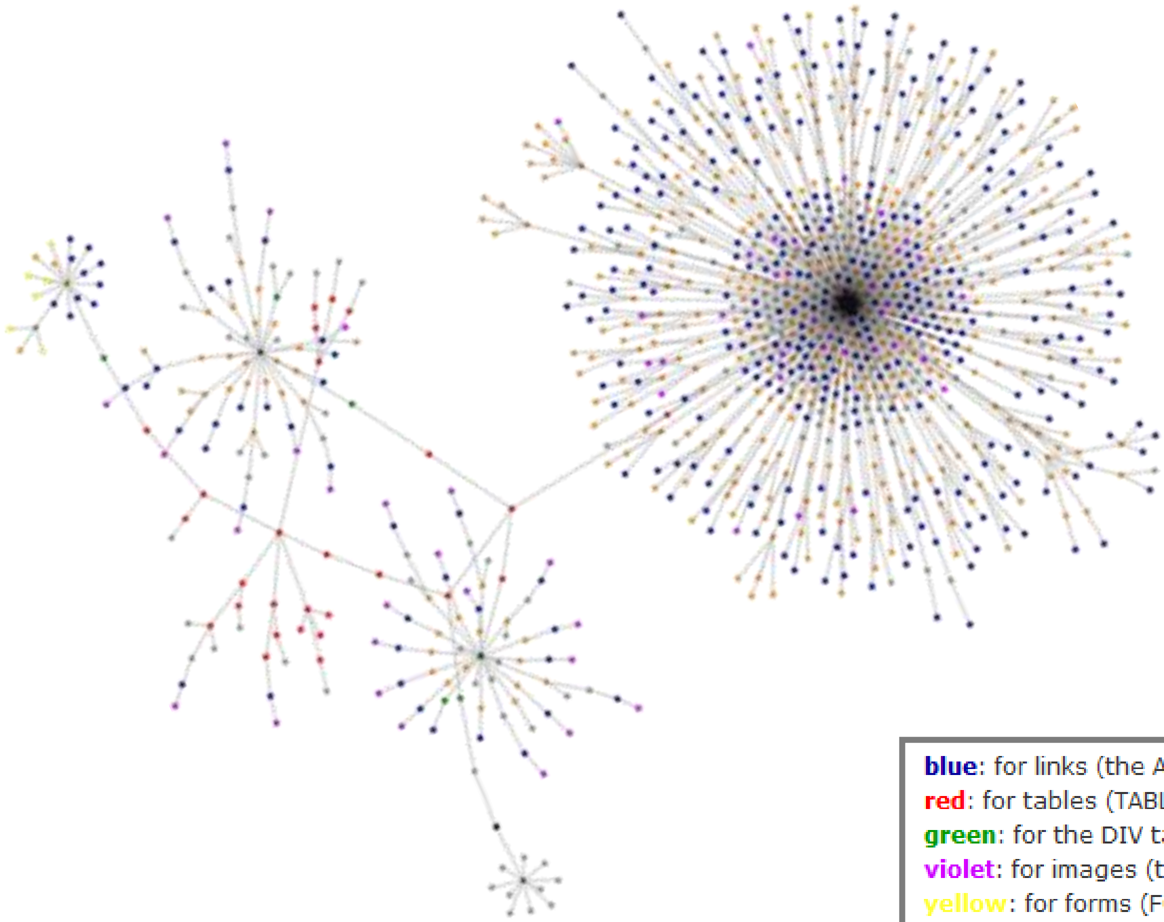


http://www.aharef.info/2006/05/websites_as_graphs.htm

cnn.com
Lots of links, lots of images.
Similar use of divs and tables for layouting purposes. (1316 tags)

blue: for links (the A tag)
red: for tables (TABLE, TR and TD tags)
green: for the DIV tag
violet: for images (the IMG tag)
yellow: for forms (FORM, INPUT, TEXTAREA, SELECT and OPTION tags)
orange: for linebreaks and blockquotes (BR, P, and BLOCKQUOTE tags)
black: the HTML tag, the root node
gray: all other tags

web pages as graphs 2



boingboing.net
One essential container that contains all other tags, essentially links , images, and tags to layout the text. A typical content driven website. (1056 tags)

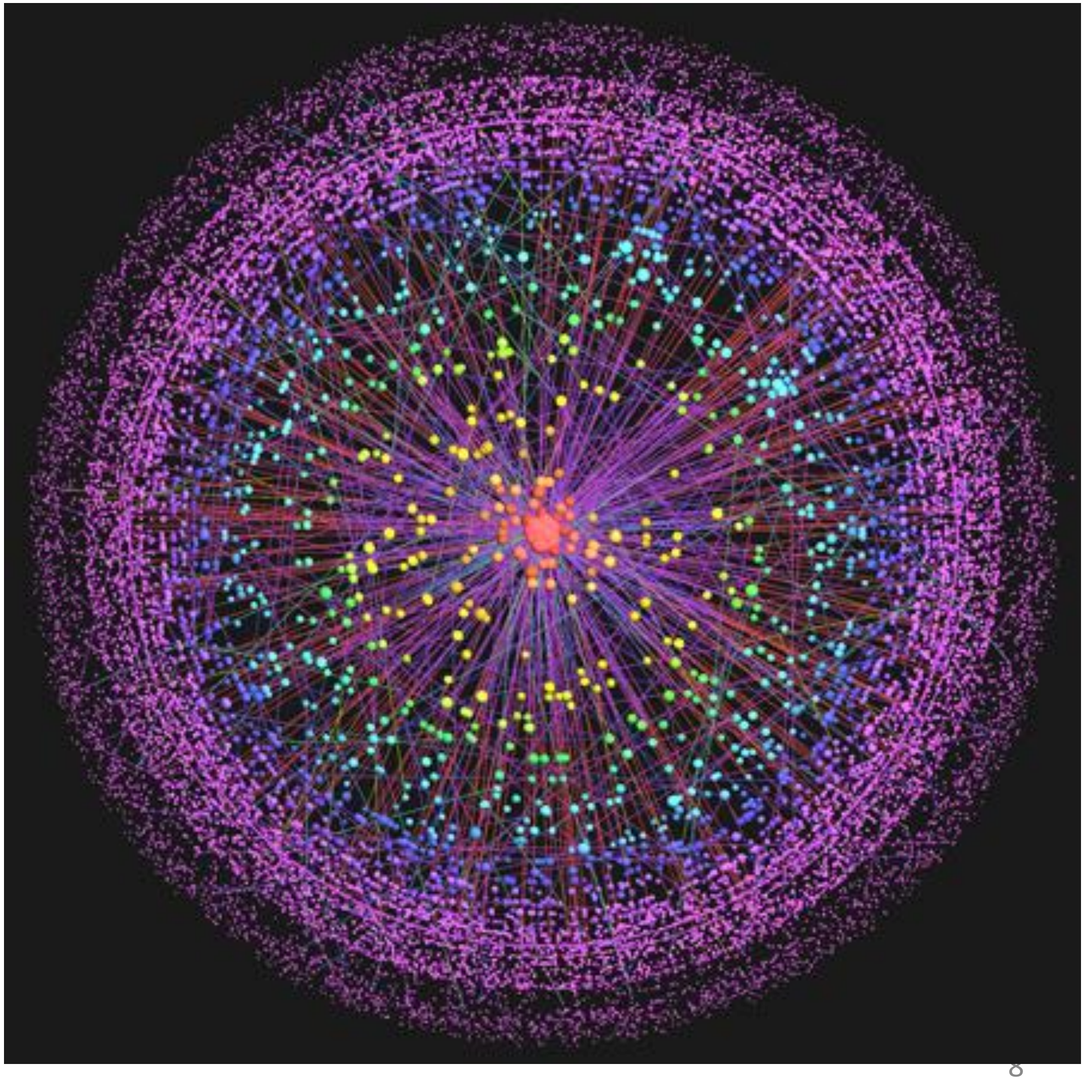
blue: for links (the A tag)
red: for tables (TABLE, TR and TD tags)
green: for the DIV tag
violet: for images (the IMG tag)
yellow: for forms (FORM, INPUT, TEXTAREA, SELECT and OPTION tags)
orange: for linebreaks and blockquotes (BR, P, and BLOCKQUOTE tags)
black: the HTML tag, the root node
gray: all other tags

http://www.aharef.info/2006/05/websites_as_graphs.htm

Internet as a graph

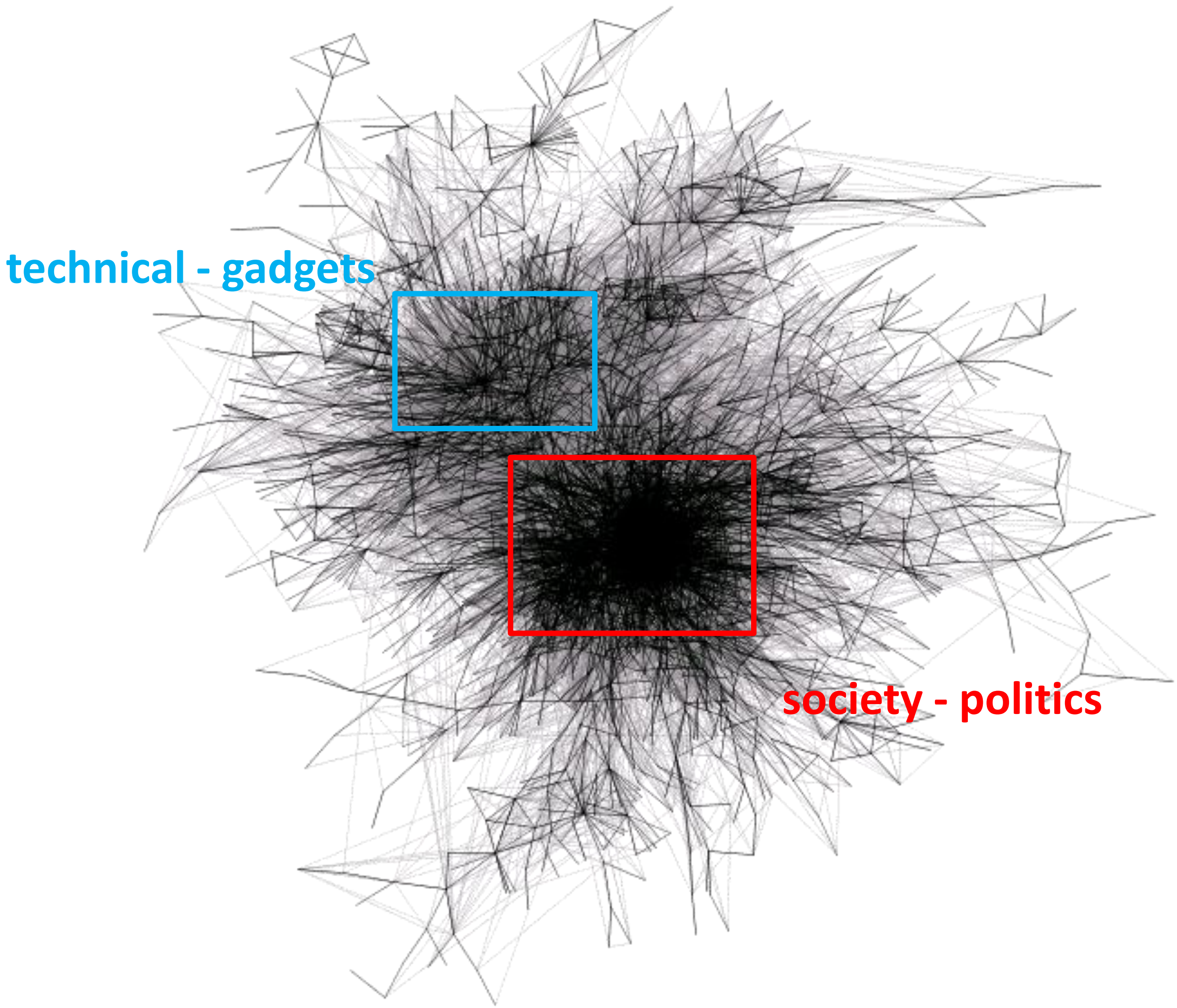
nodes = service providers
edges = connections

hierarchical structure



S. Carmi,S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir. A model of Internet topology using k-shell decomposition. PNAS 104 (27), pp. 11150-11154, 2007

blogosphere as a graph

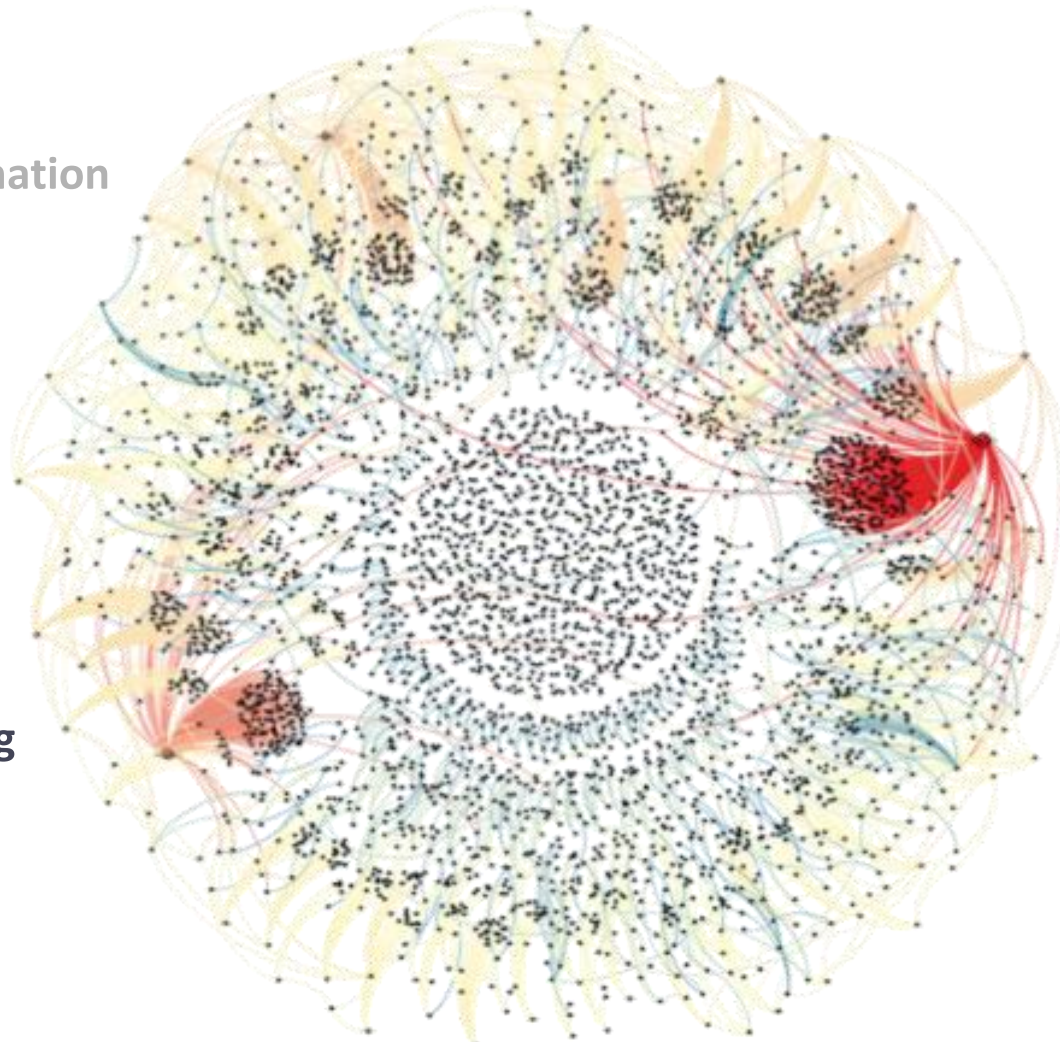


nodes = blogs
edges = hyperlinks

http://datamining.typepad.com/gallery/blog-map-gallery.html

social web as a graph

announcement of Mubarak's resignation



nodes = twitter users
edges = retweets on #jan25 hashtag

http://gephi.org/2011/the-egyptian-revolution-on-twitter/

emerging structures

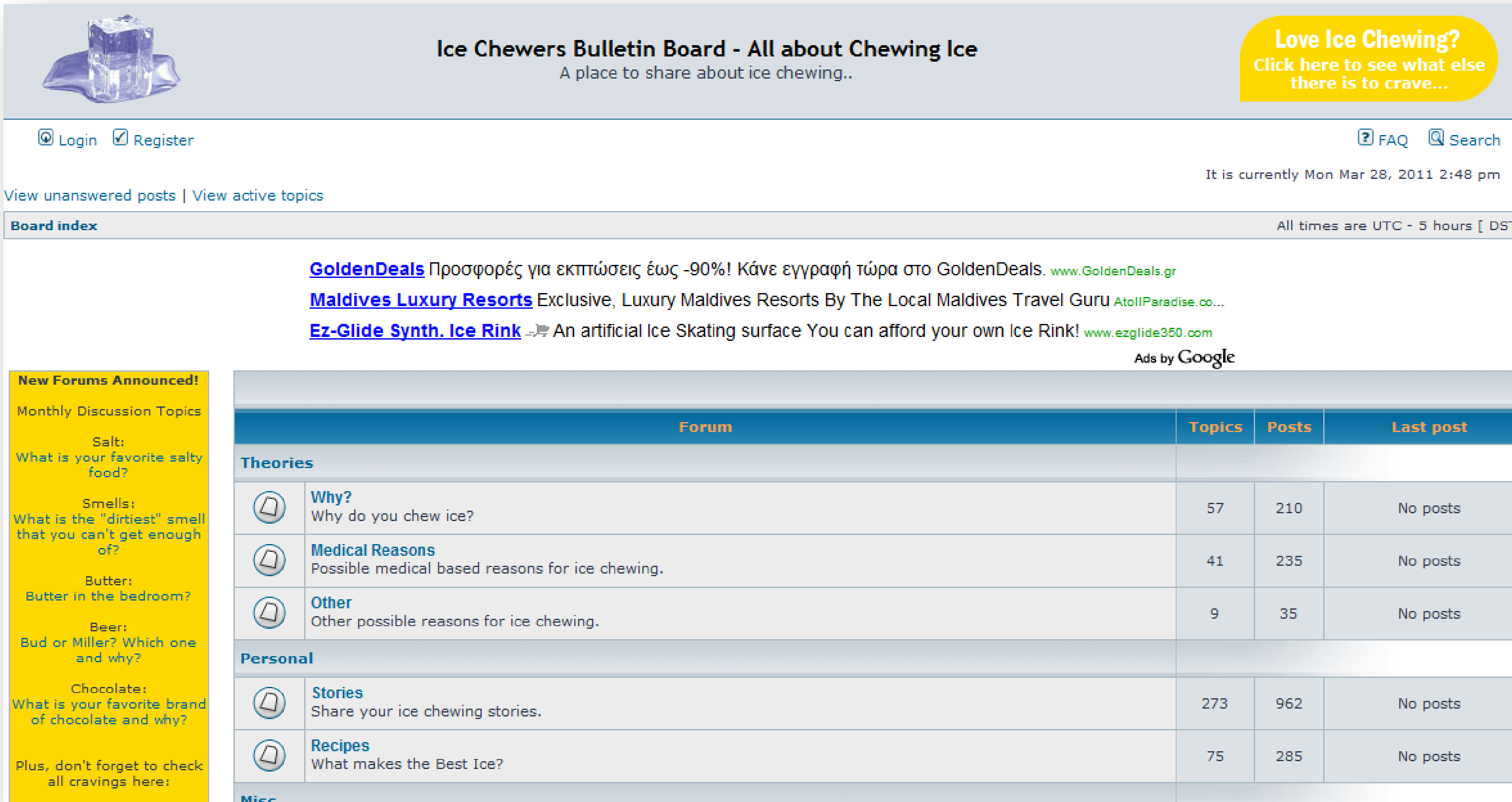
- graphs on the web present certain structural characteristics
- groups of nodes interacting with each other → dense inter-connections → functional/topical associations
- what can we gain by studying them?
  > topic analysis
  > photo clustering
  > improved recommendation methods

The concept of Community

BASICS

#1

bulleting board

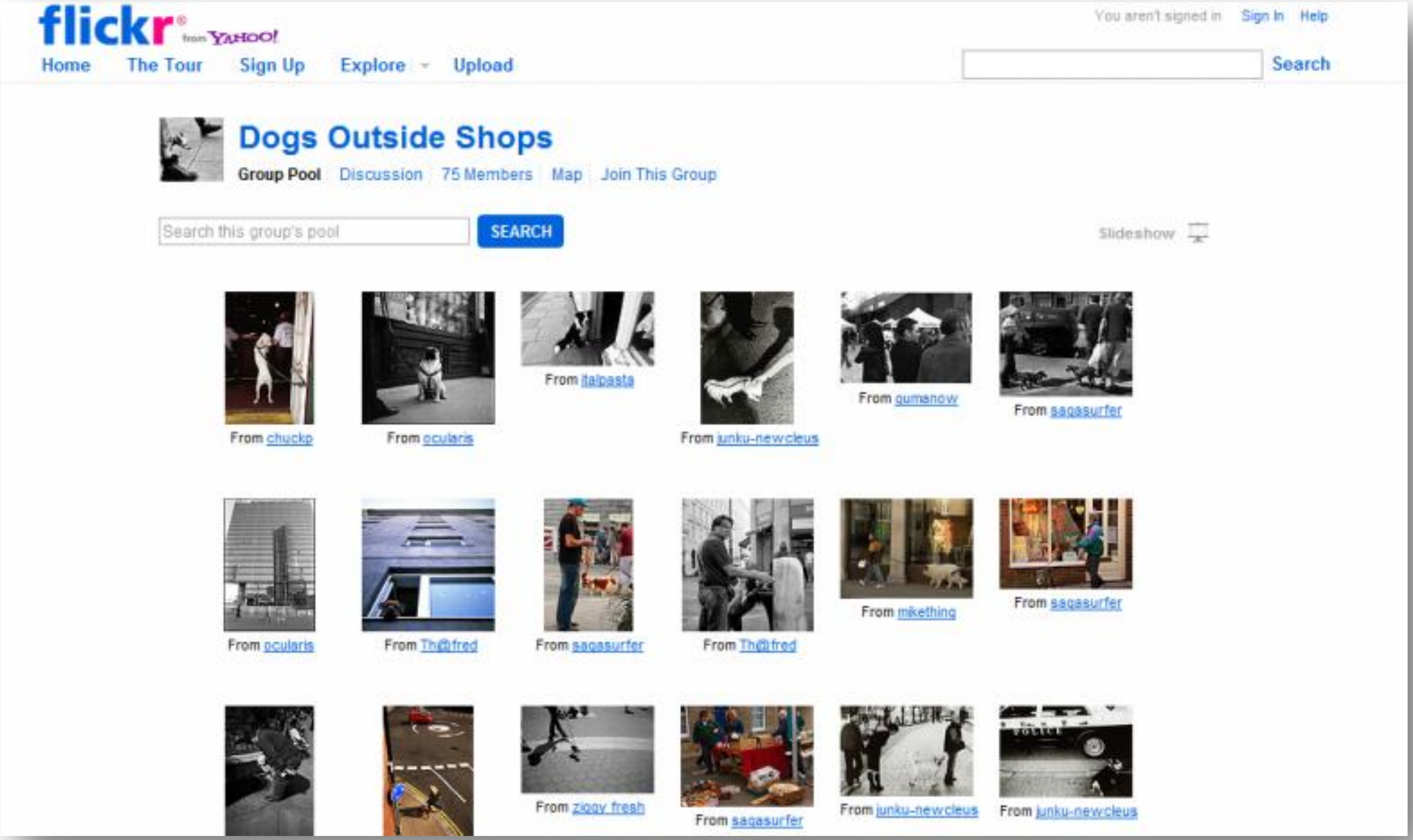


http://icechewing.iswhaticrave.com/index.php

Facebook group

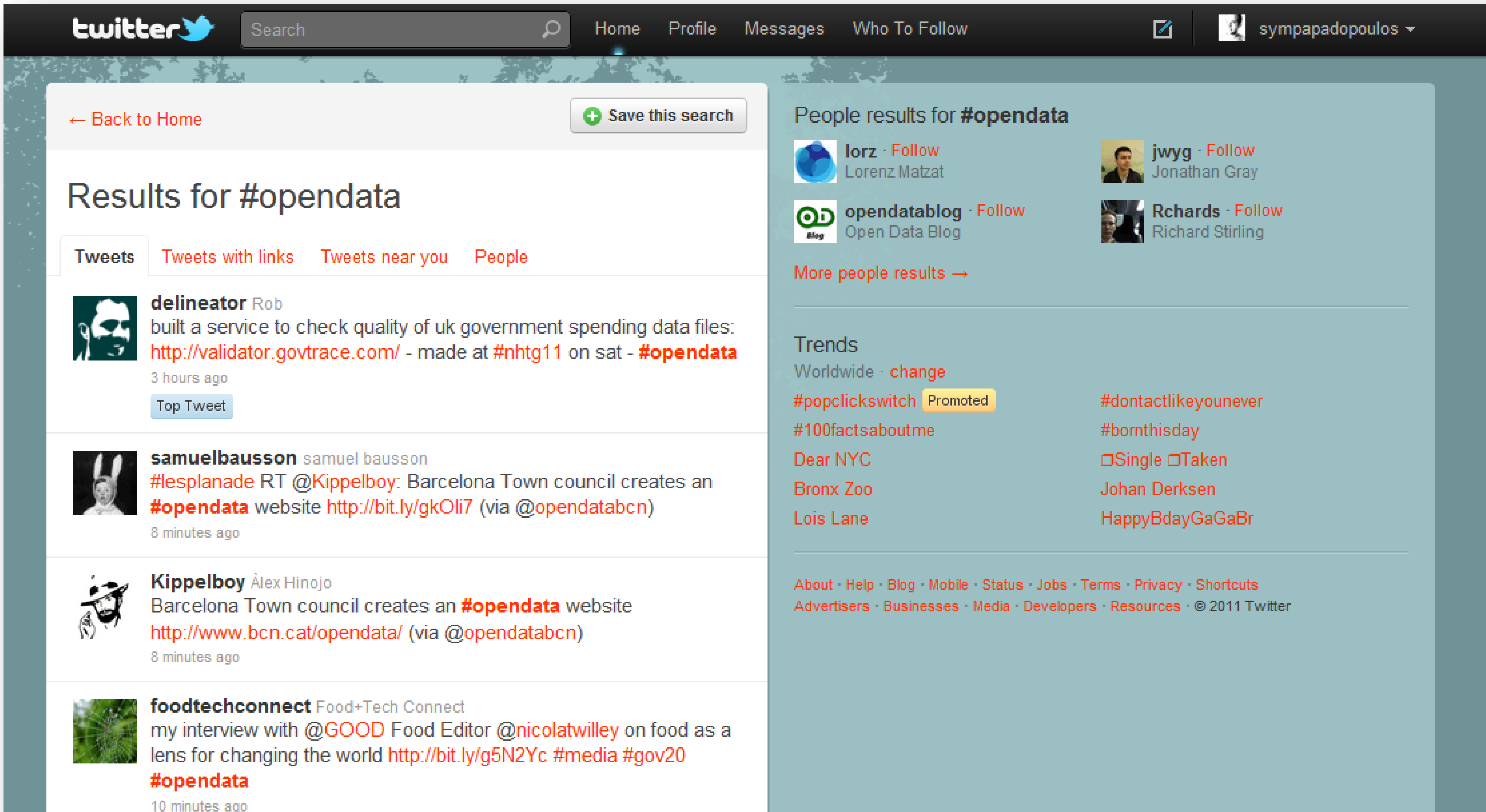
http://www.facebook.com/group.php?gid=5475162391&v=wall

Flickr group



http://www.flickr.com/groups/49246928@N00/pool/with/417646359/

#hashtag



http://twitter.com/#!/search?q=%23opendata

Web community types

- explicit
  - > the result of conscious human decision

- implicit
  - > emerging from the interactions & activities of users
  - > need special methods to be discovered

what is a Web community? *1*

Classic Web

Group of people sharing a common interest and creating web pages around it → these pages tend to link to each other more often than to the rest of the Web

R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. Trawling the web for emerging cyber-communities. Computer Networks, 31(11-16):1481-1493, 1999.

what is a Web community? 2

Web 2.0

A Web community is a set of web objects (documents and users) that includes its own semantic and logical structures.

Y. Zhang, J. Xu Yu, J. Hou. Web Communities: Analysis and Construction. Springer 2006.

what is a Web community? *3*

Complex systems

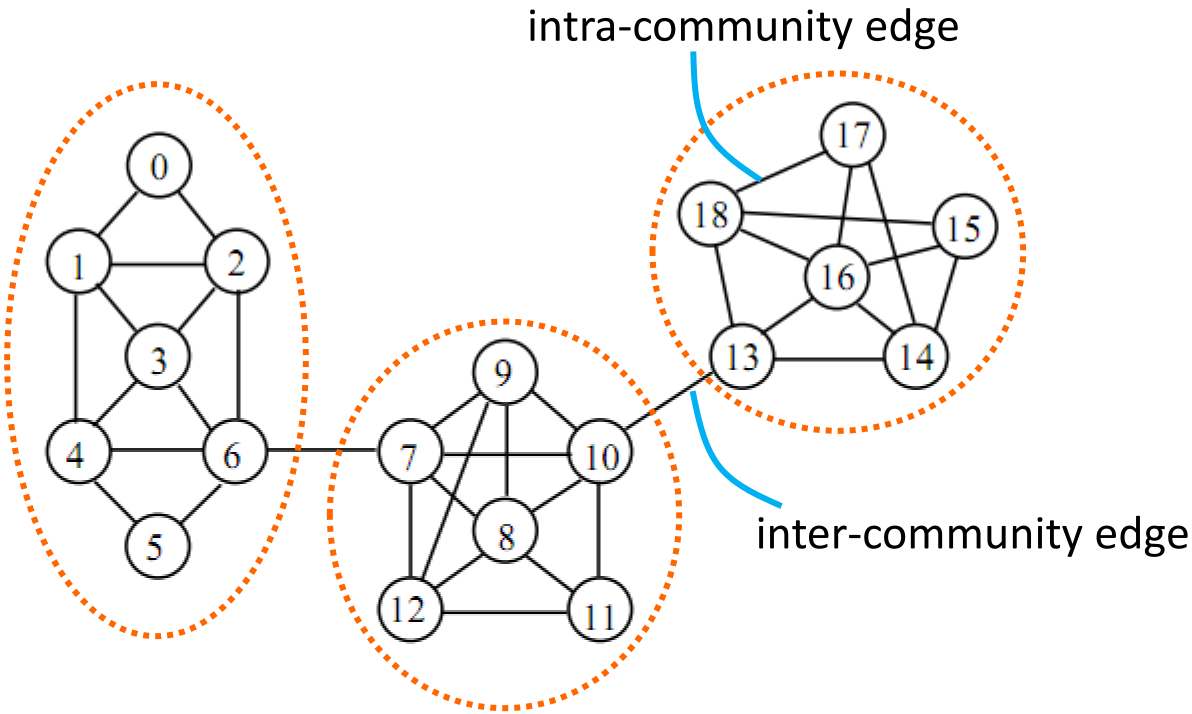
Communities correspond to groups of nodes on a graph that share common properties or have a common role in the organization/operation of the system.

S. Fortunato, C. Castellano. Community structure in graphs. arXiv:0712.2716v1, Dec 2007.

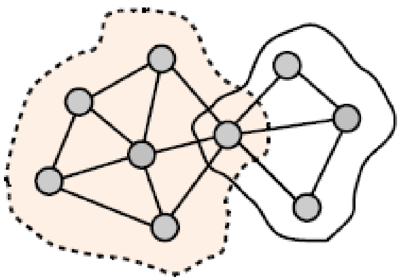
communities and graphs

- Often communities are defined with respect to a graph, ***G* = *(V,E)*** representing a set of objects (***V***) and their relations (***E***).
- Even if such graph is not explicit in the raw data, it is usually possible to construct, e.g. feature vectors → distances → thresholding → graph
- Given a graph, a community is defined as a set of nodes that are more densely connected to each other than to the rest of the network nodes.

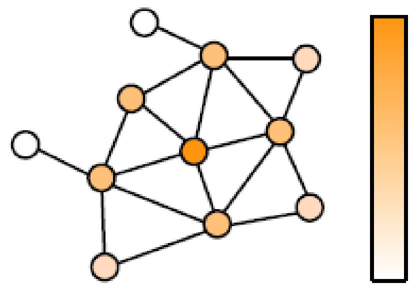
communities and graphs - example



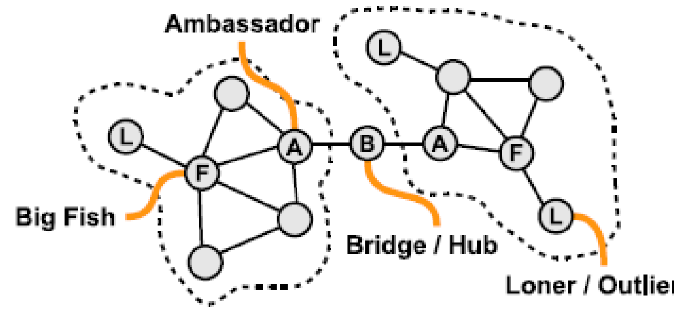
community attributes



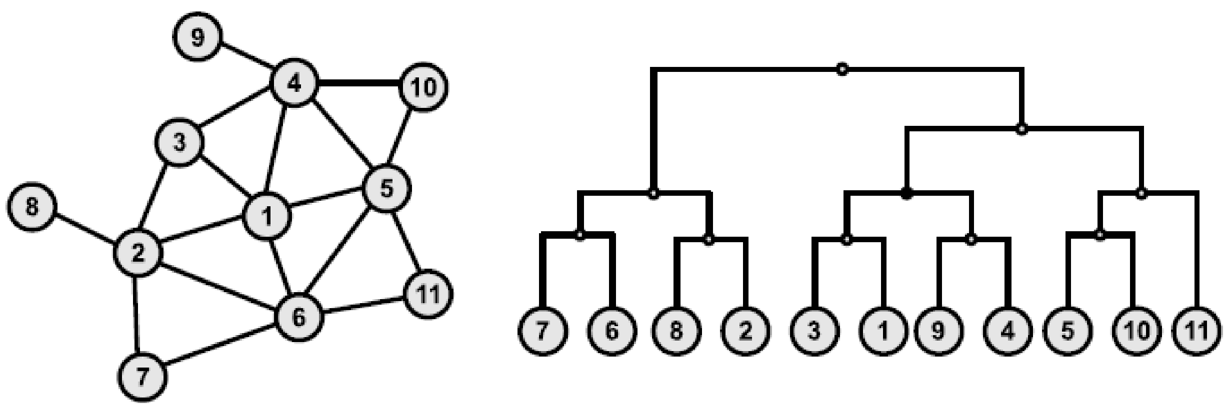
overlap



weighted participation



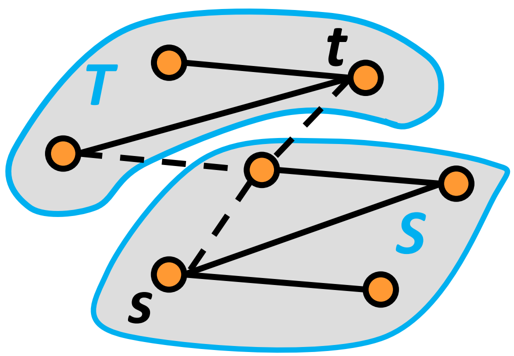
roles



hierarchy

graph cuts

- Given nodes ***u*** and ***v*** of graph ***G = (V,E)*** a **cut** is a set of edges ***C*** $\subset$ ***E***, such that the two nodes are unconnected on the graph ***G′= (V,E-C)***.



- Using ***s*** to denote a "source" node and ***t*** to denote a "terminal" node, a cut ***(S,T)*** of ***G = (V,E)*** is a partition of ***V*** in sets ***S*** and ***T = V-S***, such that ***s*** $\in$ ***S*** and ***t*** $\in$ ***T***.

modularity

- A graph can be split into communities in numerous ways, i.e. for each graph there are many possible community structures. In the simple case, a community structure is defined as a graph partition into a set of node sets $\boldsymbol{C = \{C_i\}}$.

- To provide a measure of the quality of a community structure, we make use of **modularity**.

- Modularity quantifies the extent to which a given graph partition into communities presents a systematic tendency to have more intra-community links than the same community structure would present if the links would be rewired under ER (Erdos-Renyi) graph model.

modularity computation

- Modularity is computed as follows:

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j)$$

Observed number of intra-community edges.

Expected number of edges between ***i*** and ***j***, if edges are placed randomly.

> ***A_{ij}*:** adjacency matrix

> ***k_i*:** degree of node ***i***

> ***c_i***: community of node ***i***

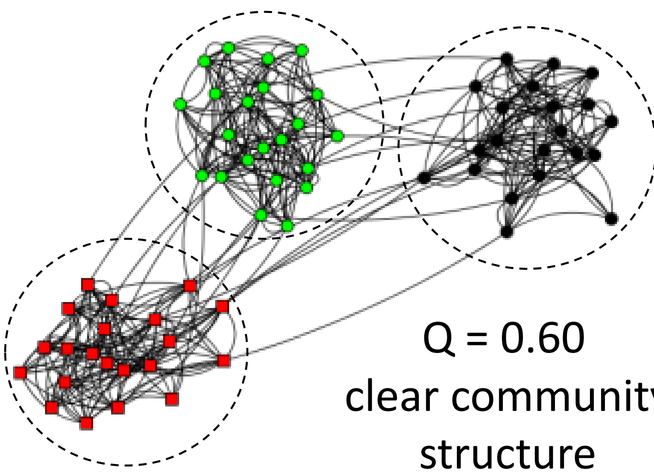
> ***$\delta(c_i,c_j) = 1$*** if ***i*, *j*** belong to the same community

> ***m***: number of edges on the graph

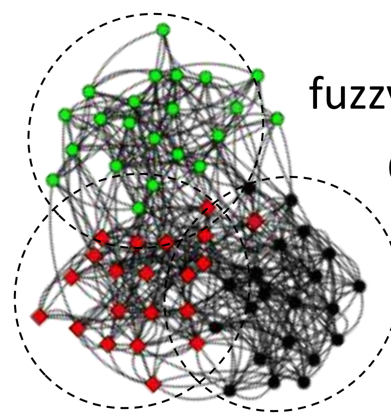
modularity - example

- In a random graph (ER model), we expect that any possible partition would lead to ***Q = 0***.

- Typically, in non-random graphs modularity takes values between 0.3 and 0.7.



Q = 0.60
clear community structure



fuzzy communities
Q = 0.37

Taxonomy of Community Detection Algorithms

METHODS

#2

graph partition

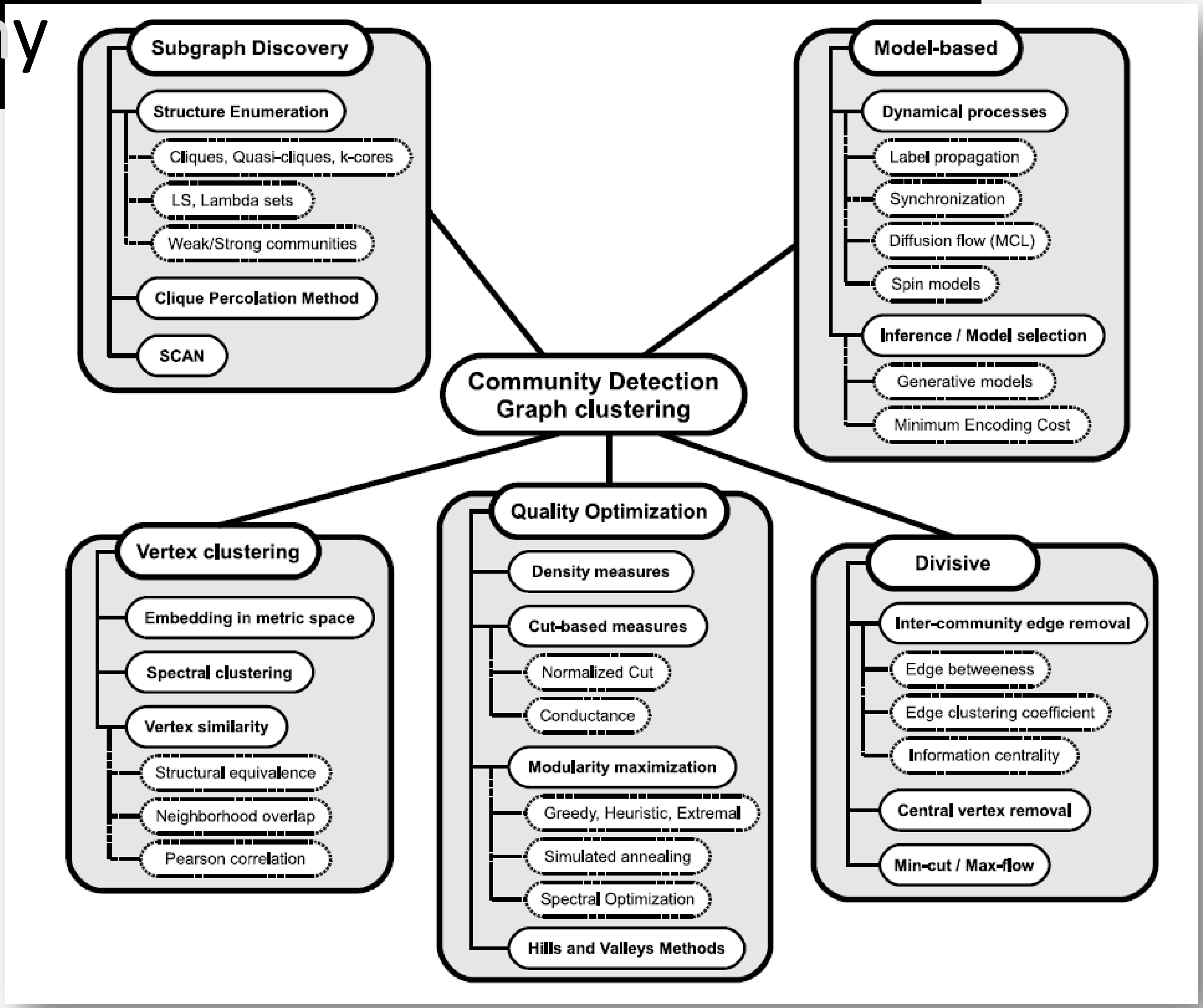
- Given a graph ***G=(V,E)***, find a partition of ***V*** in ***k*** disjoint subsets, such that the number of edges in ***E*** of which the endpoints belong to different subsets is minimized.
- Various solutions: Kernighan-Lin algorithm [Kernighan70], spectral bisection [Pothen90].
- Multi-level partition (metis) [Karypis99]: Repeated application of bisection until the graph is partitioned into ***k*** parts under constraint to the sizes of the subsets.
- Not satisfactory solution, since the number of communities needs to be provided as input to the algorithm. Sometimes event the community sizes need to be provided as inputs.

B. W. Kernighan, S. Lin. An Efficient Heuristic Procedure for Partitioning of Electrical Circuits. Bell Systems Technical Journal, Vol. 49, No. 2, pp. 291- 307, February 1970.

A. Pothen, H.D. Simon and K.-P. Liou. Partitioning sparse matrices with eigenvectors of graphs. SIAM journal of Matrix Analysis and Applications, 11: 430-452, 1990.

G. Karypis and V. Kumar, A fast and high quality multilevel scheme for partitioning irregular graphs, SIAM J. Sci. Comput. 20 (1): 359–392, 1999.

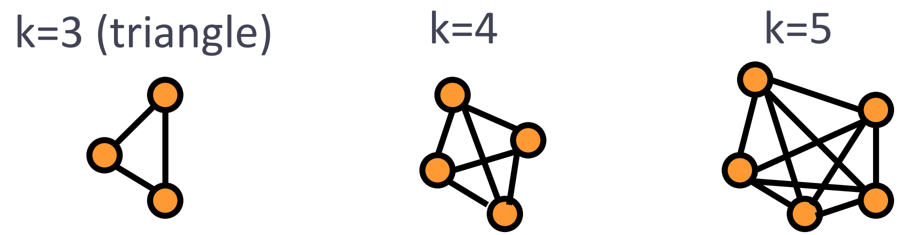
taxonomy



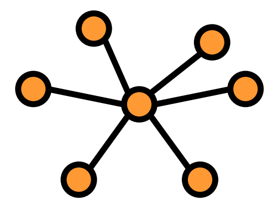
S. Papadopoulos, Y. Kompatsiaris, A. Vakali, P. Spyridonos. “Community detection in Social Media”. In Data Mining and Knowledge Discovery, Springer (accepted 2011)

subgraph discovery *1*

- k-clique

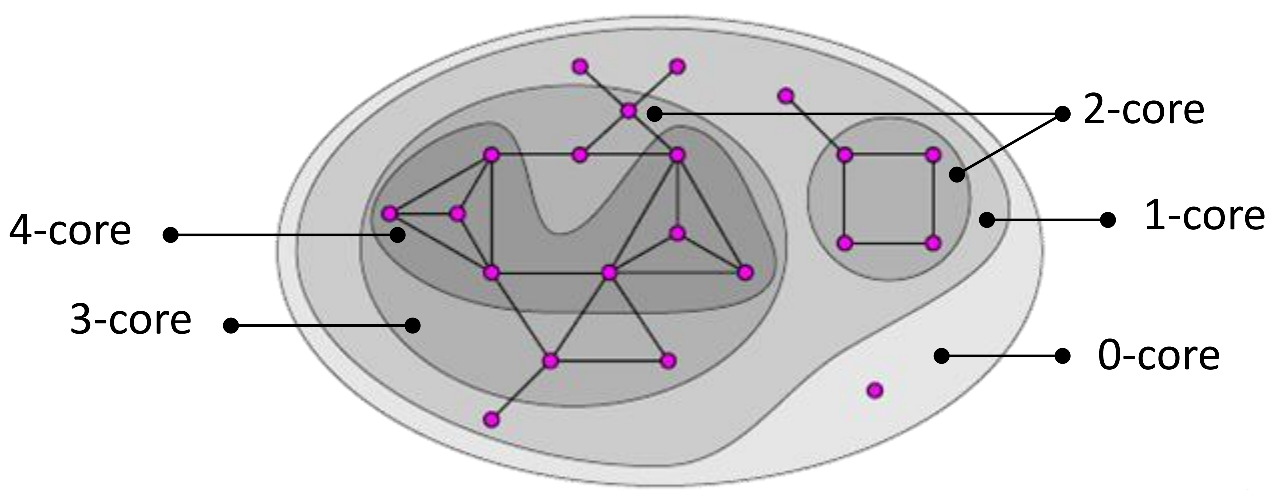


- N-clique



N=2 (star)

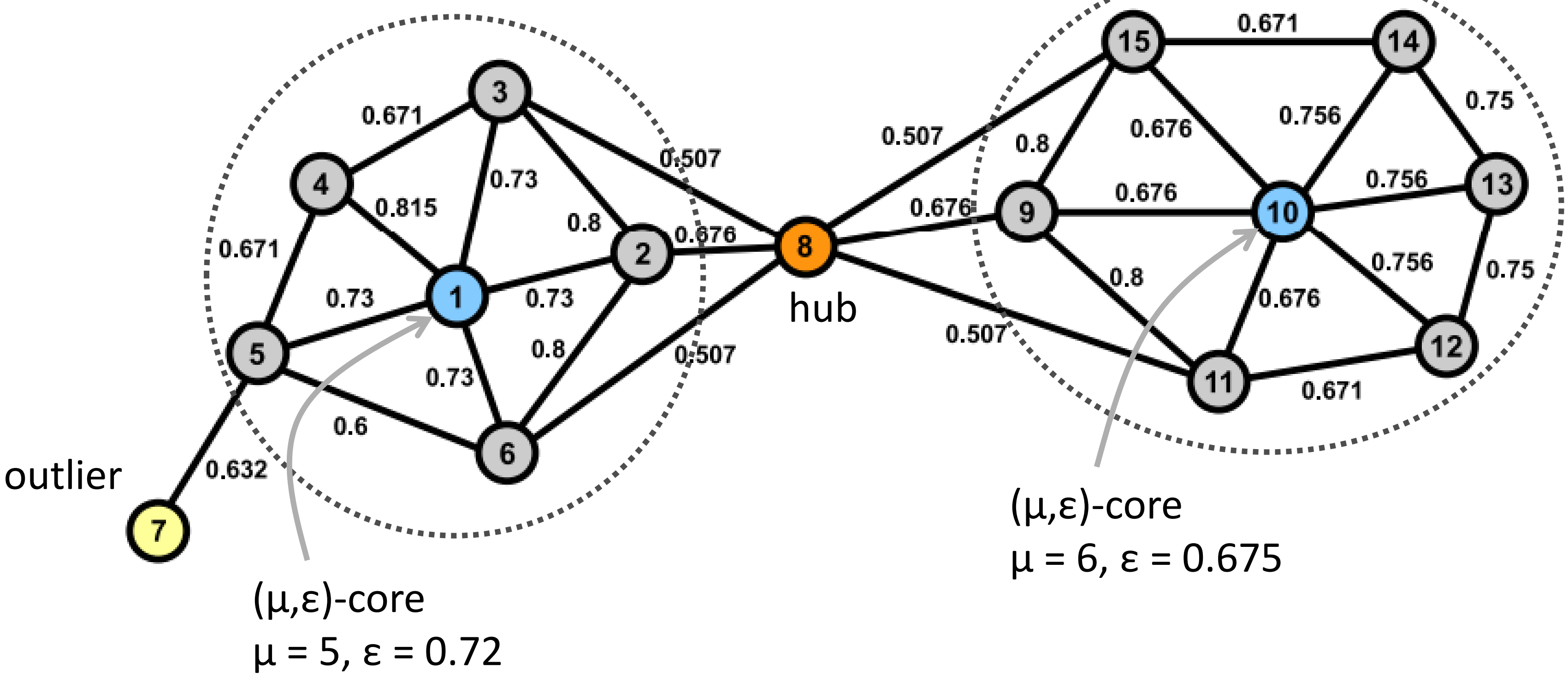
- k-core



subgraph discovery 2

- (μ,ε)-core:
  > based on the concept of structural similarity

$$\sigma(v, w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{\sqrt{|\Gamma(v)| \cdot |\Gamma(w)|}}$$



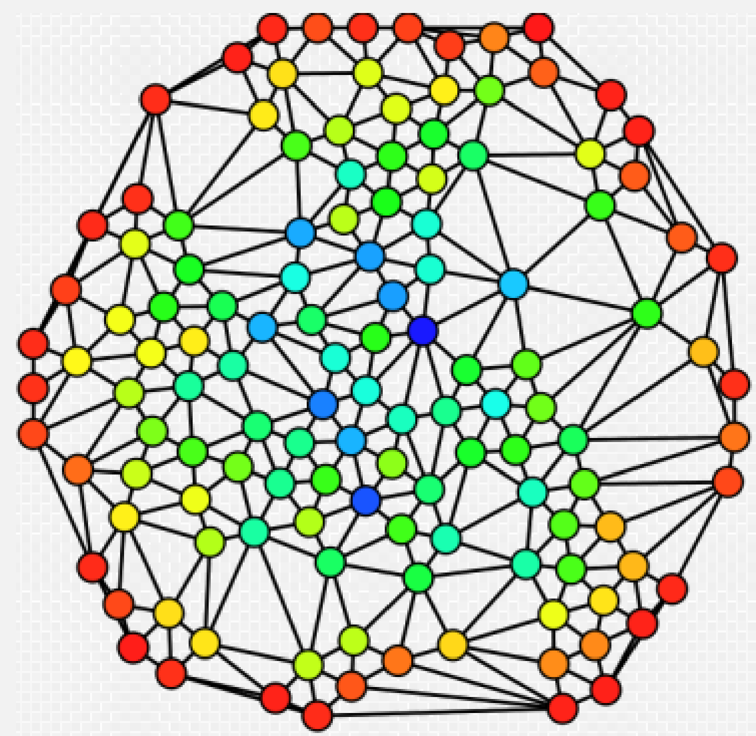
divisive - use of edge centrality

- Find edges that stand between communities.
- Progressively remove more “central” edges until the graph breaks into separate communities.
- As the graph splitting progresses, new communities are assigned to a hierarchical structure.
- Edge centrality is defined similarly to node centrality:

$$bc(v) = \sum_{\substack{s \neq t \neq v \\ s,t \in V}} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$$

$\sigma_{s,t}(v)$: number of paths from node ***s*** to ***t*** that include node ***v***

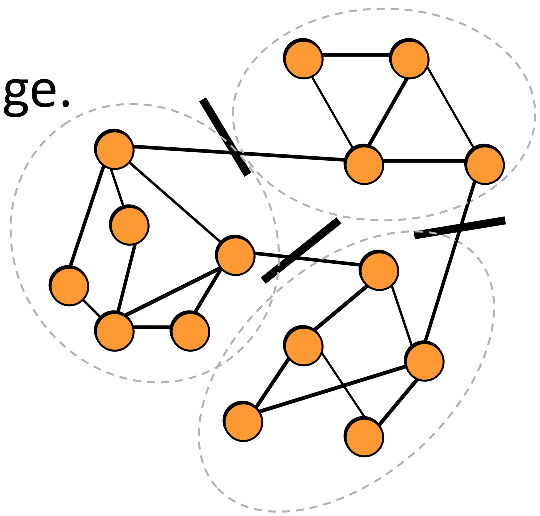
$\sigma_{s,t}$: total number of paths from ***s*** to ***t***



Depiction of node centrality:
red (min) → blue (max)

Girvan - Newman algorithm

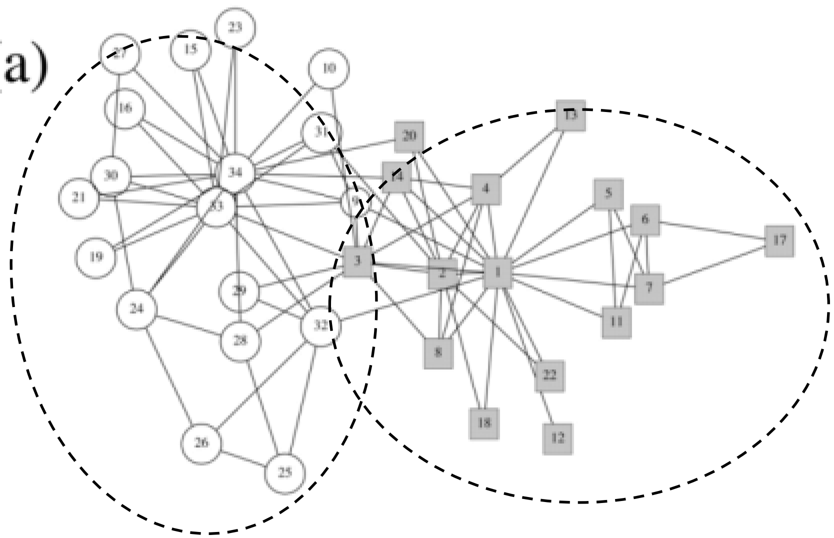
- GN algorithm is one of the most important algorithms stimulating a whole wave of community detection methods.
- Basic principle:
  > Compute betweenness centrality for each edge.
  > Remove edge with highest score.
  > Re-compute all scores.
  > Repeat 2nd step.
- Complexity: ***O(n³)***
- Many variations have been presented to improve precision by use of different betweenness measures or reduce complexity, e.g. by sampling or local computations.



Girvan, M., Newman, M.E.J. “Community structure in social and biological networks”. In Proceedings of National Academy of Science, U. S. A. 99(12), 7821–7826, 2002

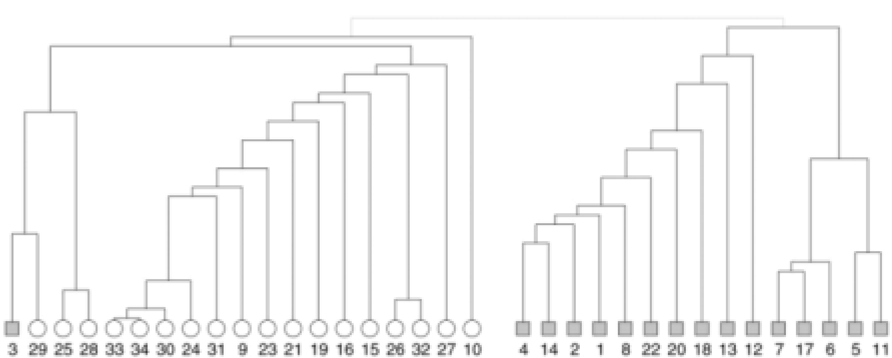
Girvan - Newman (example)

(a)



Social network in Zachary karate club

(b)

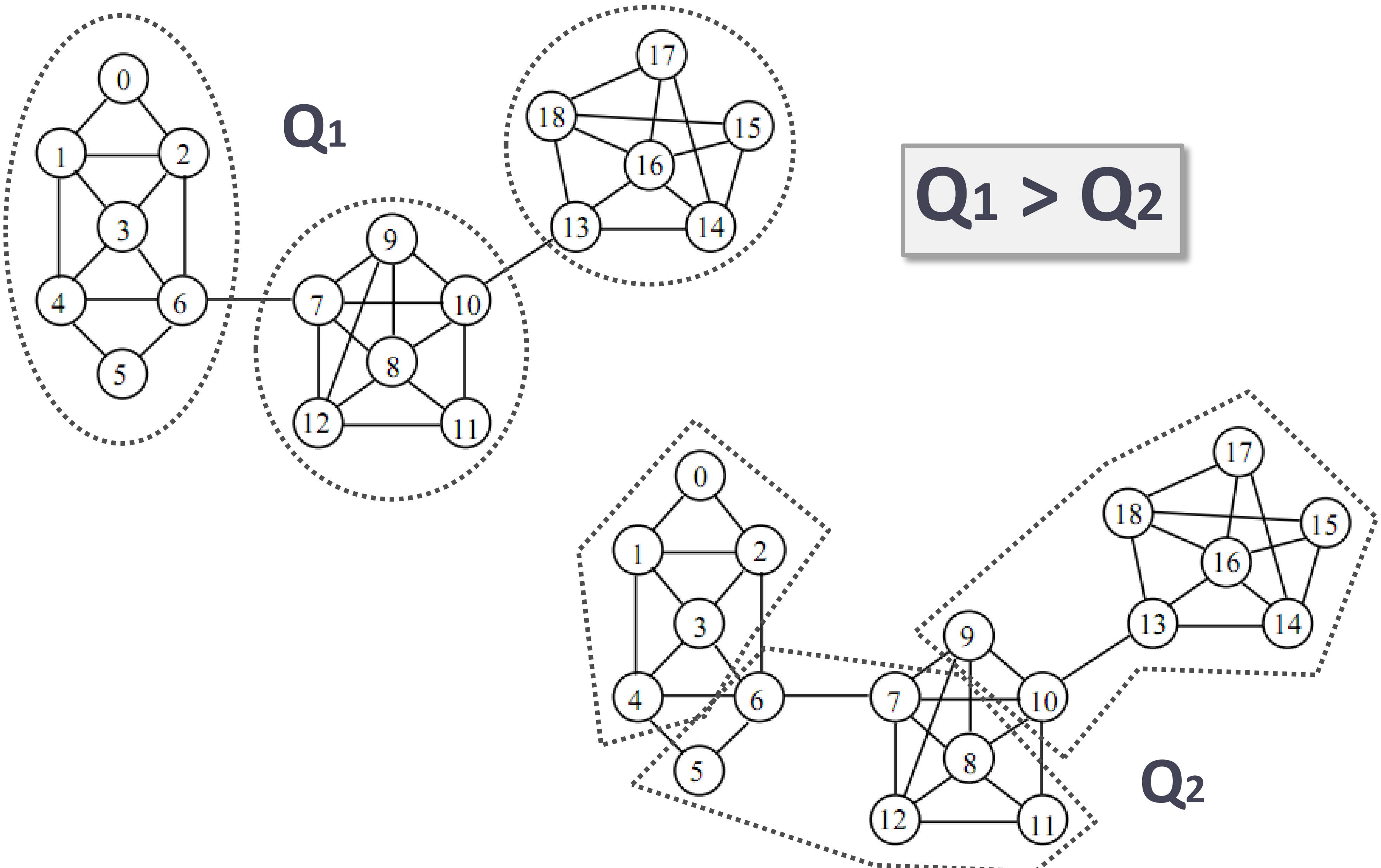


Hierarchical community structure detected by the algorithm.

modularity maximization

- Modularity indicates the quality of a given community structure.
- A class of methods seeks for a community structure that maximizes the value of modularity.
- The search space is exponential with respect to the number of nodes, thus approximate and heuristic schemes are devised.

modularity maximization example



$Q_1 > Q_2$

greedy solution

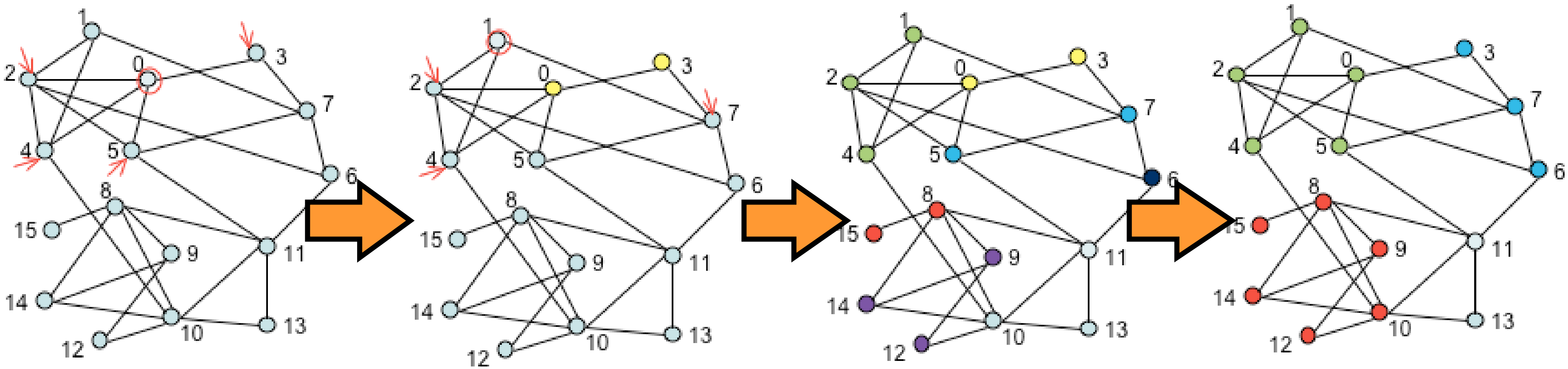
- Merge nodes trying in each merging step to maximize the graph modularity (Newman, 2004).
- Leads to hierarchical structure.
- Complexity in a sparse graph: $\boldsymbol{O(n^2)}$
- Use of appropriate data structures (max-heaps) and heuristics can lead to complexity reduction (Clauset et al., 2004) → $\boldsymbol{O(n\log^2 n)}$ but can also lead to the formation of gigantic communities.

M. E. J. Newman. Fast algorithm for detecting community structure in networks. Physical Review E, 69:066133, 2004

A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. Physical Review E, vol. 70, Issue 6, id. 066111, 2004.

efficient solution 1

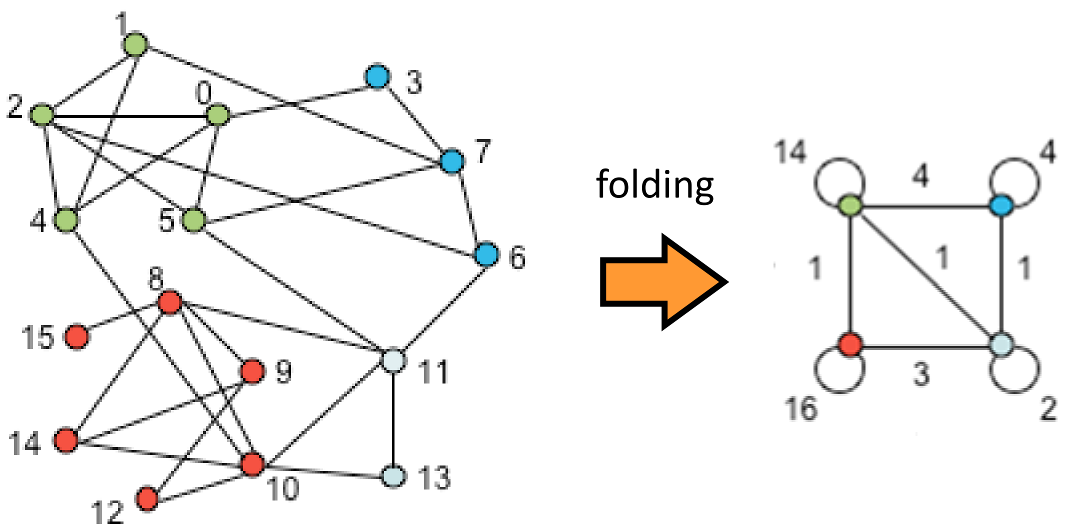
V. D. Blondel, J. Guillaume, R. Lambiotte and E. Lefebvre: Fast unfolding of communities in large networks, J. Stat. Mech., 2008.



After a single pass. After four passes

- Initially, each node belongs to its own community (N nodes → N communities)
- We go through each node with a standard order. To each node, we assign the community of their neighbor as long as this leads to an increase in modularity.
- This step is repeated many times until a local modularity maximum is found.

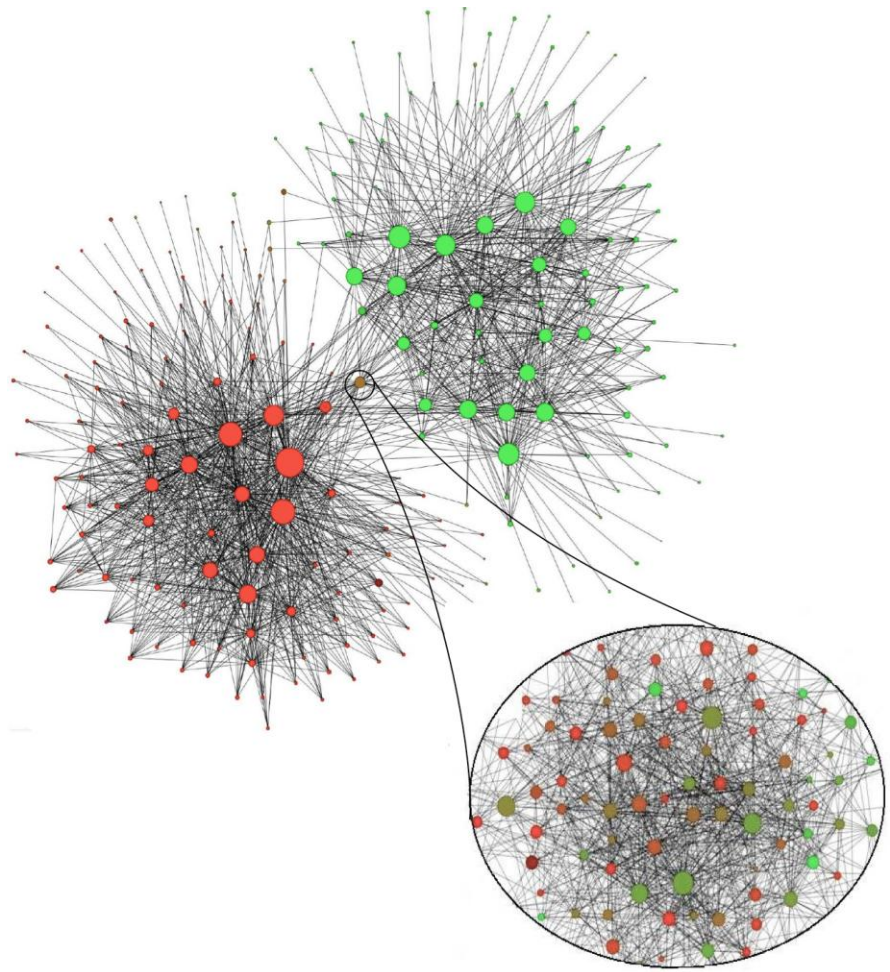
efficient solution 2



- **Folding:** Create new graph in which nodes correspond to the communities detected in the previous step.
- Edge weights between community nodes are defined by the number of inter-community edges.
- Folding ensures rapid decrease in the number of nodes that need to be examined and thus enables large-scale application of the method.

efficient solution - example

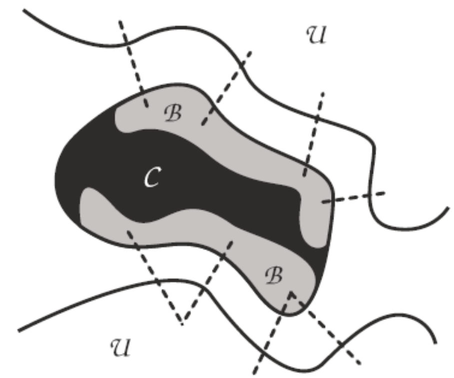
- Application on a graph with 118 million nodes and one billion edges in 152 minutes.
- Application on a mobile phone call network (snapshot).
- Hierarchical community structure.



local methods *1*

- More often than not we cannot have access to the whole graph (e.g. Web).

U: unexplored portion of graph



B: community boundary

C: “inside” part of community

- In that case local methods are valuable for detecting communities.

- A local method starts from a set of nodes (seeds) and expands the community boundaries using some criterion to stop.

local methods 2

- Local modularity

$$R = \frac{\sum_{ij} B_{ij}\delta(i,j)}{\sum_{ij} B_{ij}}$$

$$B_{ij} = \begin{cases} 1 & \text{if vertices } i \text{ and } j \text{ are connected,} \\ & \text{and either vertex is in } \mathcal{B} \\ 0 & \text{otherwise.} \end{cases}$$

A. Clauset. Finding local community structure in networks. Physical Review E, 72, 026132, 2005.

- Subgraph modularity

$$M = \frac{ind(S)}{outd(S)}$$

ind(S): in-degree of subgraph ***S***
outd(S): out-degree of subgraph ***S***

F. Luo, J. Z. Wang, E. Promislow. Exploring Local Community Structures in Large Networks. In Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence (December, 2006). Web Intelligence. IEEE Computer Society, Washington, DC, 233-239, 2006.

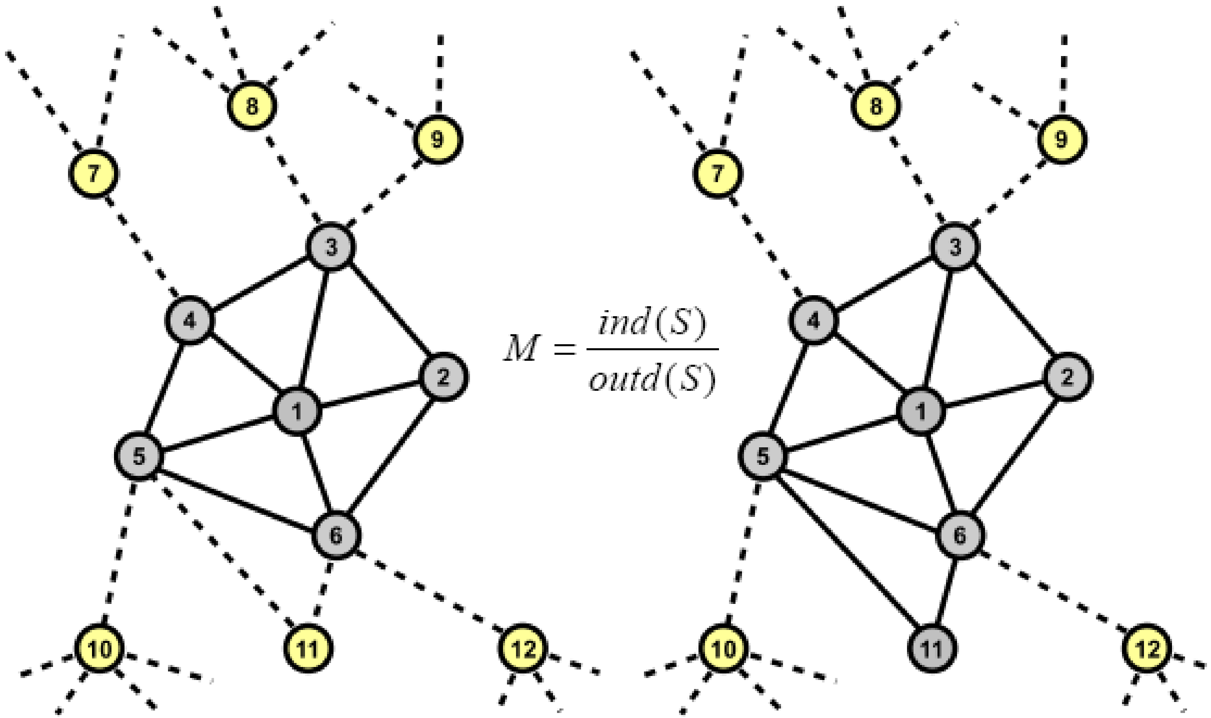
- Node outwardness

$$\Omega_v(C) = \frac{1}{k_v} \sum_{i \in n(v)} \Big([i \notin C] - [i \in C] \Big)$$

k_u: degree of ***u***
n(u): neighborhood of ***i***

J. P. Bagrow. Evaluating Local Community Methods in Networks. J. Stat. Mech., 2008 (5): P05001.

local methods - example



$$M = \frac{ind(S)}{outd(S)}$$

(a) $M(S) = 1.429$ (b) $M(S_{exp}) = 2.4$

F. Luo, J. Z. Wang, E. Promislow. Exploring Local Community Structures in Large Networks. In Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence (December, 2006). Web Intelligence. IEEE Computer Society, Washington, DC, 233-239, 2006.

evaluation

- How to quantify how well a given method detects communities that exist in a graph
  - > Synthetic graph with known (generated) community structure.
  - > Small real-world graphs with known community structure.
  - > Large graphs and subjective discussion/evaluation of results.
  - > Deployment of community structure in some task (e.g. tag recommendation) and evaluation of results comparing the performance against a community structure produced by a baseline/state-of-the-art method or against a solution that does not leverage community structure.

- Computational complexity

- Memory requirements

compare to ground truth

- Fraction of correctly assigned nodes
  - not always well-defined, e.g. in the case where two community structures have different numbers of communities

- Normalized Mutual Information (NMI)

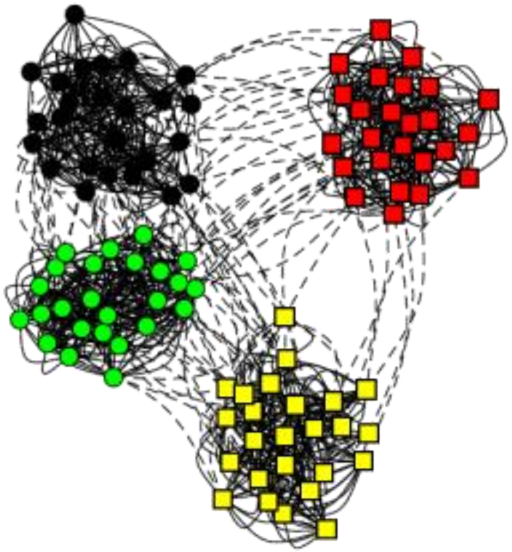
$$NMI(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij} log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{c_A} N_{i.} log\left(\frac{N_{i.}}{N}\right) + \sum_{j=1}^{c_B} N_{.j} log\left(\frac{N_{.j}}{N}\right)}$$

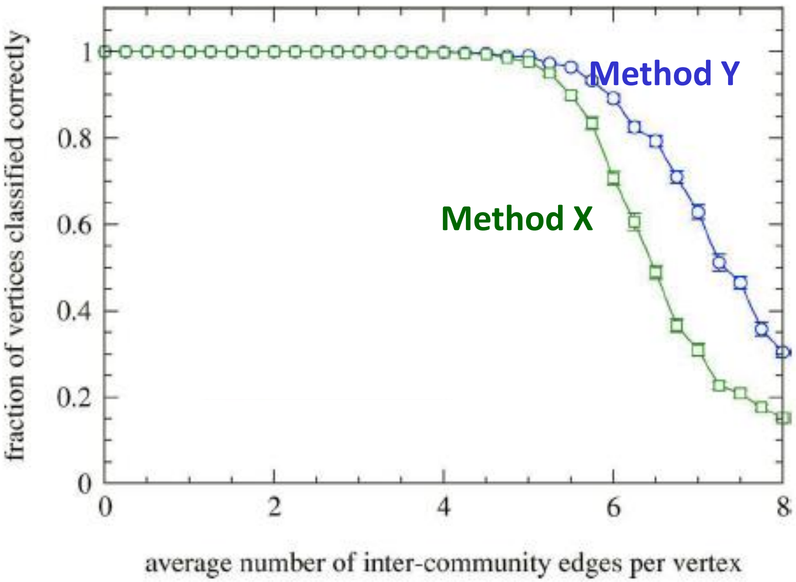
  - N_{ij}: confusion matrix
  - N_i: sum over elements of row i
  - c_A, c_B: number of communities in “true” and “detected” community structure

synthetic graphs *1*

- The simplest synthetic test:
  - > A graph of 128 nodes split in 4 communities of 32 members. All nodes have the same degree. Inter-community edges are randomly placed (parameter z_{out}).

M. E. J. Newman, M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69:026113, 2004.

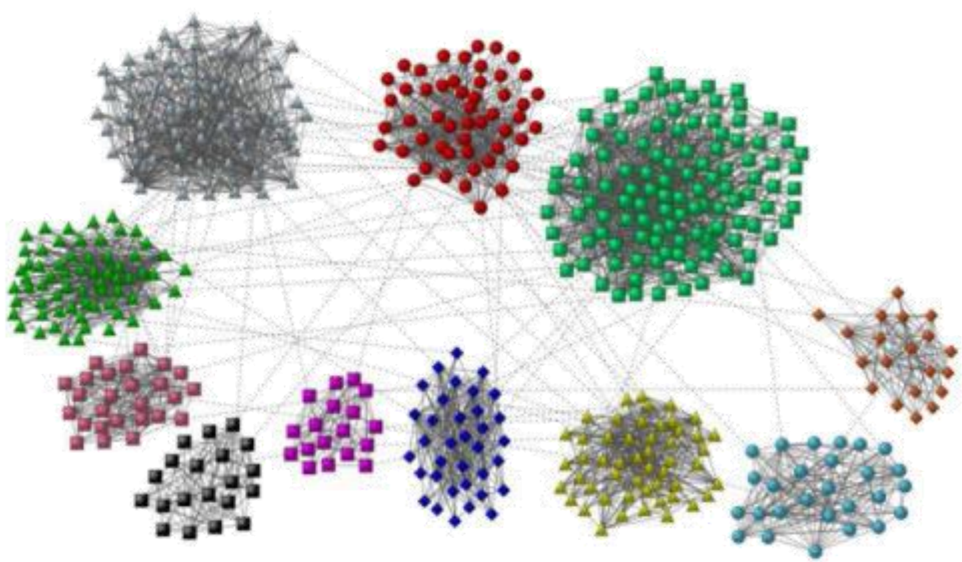




synthetic graphs 2

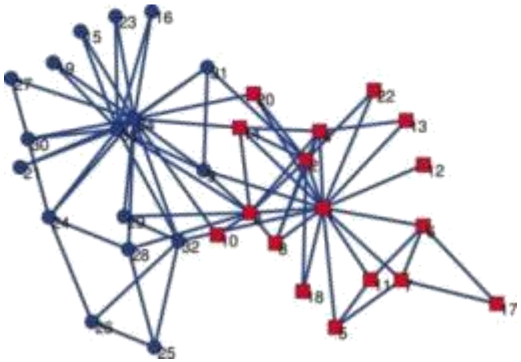
- The previous synthetic test is very simple. → Most algorithms do good.
  - Harder test in which the community sizes and node degrees follow a power law distribution.

A. Lancichinetti, S. Fortunato, J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 2009



small real-world graphs

- Zachary karate club
- American football teams
- Dolphin social network
- Researchers & conferences (DBLP)



evaluation in the context of task

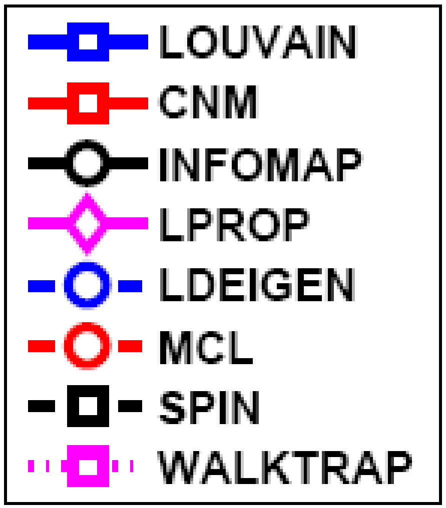
- We exploit community structure detected by a given algorithm for some information retrieval task, e.g. tag recommendation.
- We compare between the following:
  - > one using the community structure produced by a reference method (baseline)
  - > one using the community structure detected by our method
- We collect usage data (e.g. tagging activities) that are publicly available through a web service (e.g. flickr).
- We hold a portion of the data for training (community detection) and the rest we use as ground truth.
- We compare between the two methods by means of information retrieval measures (precision, recall).

performance evaluation - theoretical

Method	Complex-A	Complex-B	Scale
	Cohesive substructure detection		
Bron-Kerbosch [15]	$O(3^{n/3})$	$O(3^{n/3})$	S
k-core detection (Batagelj and Zaversnik [10])	$O(n^2)$	$O(n)$	L
SCAN (Xu et al. [110])	$O(n^2)$	$O(n)$	L
	Vertex clustering		
embedding in space + k-means	$O(C\,n^2)$	$O(C\,n^2)$	M
Walktrap (Pons and Latapy [85])	$O(n^4)$	$O(n^2 \log n)$	M
Donetti-Munoz [29]	$O(C\,n^2)$	$O(C\,n^2)$	M
	Community quality optimization		
Clauset-Newman-Moore [23]	$O(n^2\,d\,\log n)$	$O(n\,\log^2 n)$	M
Extremal optimization (Arenas et al. [4])	$O(n^2 \log n)$	$O(n^2 \log n)$	M
Spectral optimization (Newman [77])	$O(n^3\,d)$	$O(n^2 \log n)$	M
Community folding (Blondel et al. [12])	$O(n^2)$	$O(n)$	L
	Divisive		
Girvan-Newman [43]	$O(n^5)$	$O(n^3)$	S
Information centrality (Fortunato et al. [35])	$O(n^7)$	$O(n^4)$	S
Edge clustering coefficient (Radicchi et al. [88])	$O(n^6)$	$O(n^2)$	M
Max flow + Gomory-Hu tree (Ino et al. [49])	$O(n^4\,\log n)$	$O(n^3 \log n)$	S
	Model-based		
MCL (Van Dongen [30])	$O(n^3)$	$O(n^3)$	M
Minimum encoding cost (Chakrabarti [18])	$O(n^2)$	$O(n)$	L
Label propagation (Raghavan [89], Leung [62])	$O(n^2)$	$O(n)$	L
Infomap (Rosvall and Bergstrom [92])	$O(n^2 \log n)$	$O(n \log n)$	L

S. Papadopoulos, Y. Kompatsiaris, A. Vakali, P. Spyridonos. “Community detection in Social Media”. In Data Mining and Knowledge Discovery, Springer (accepted 2011)

performance evaluation - empirical *1*



LOUVAIN

Modularity maximization with the use of heuristics and community folding.

V. D. Blondel, J. Guillaume, R. Lambiotte and E. Lefebvre: Fast unfolding of communities in large networks, J. Stat. Mech., 2008.

CNM

Modularity maximization using max-heaps and hierarchical merging.

A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. Phys. Rev. E70 (6), 066111, 2004.

INFOMAP

Description compression (Huffman code) of information flow in the graph.

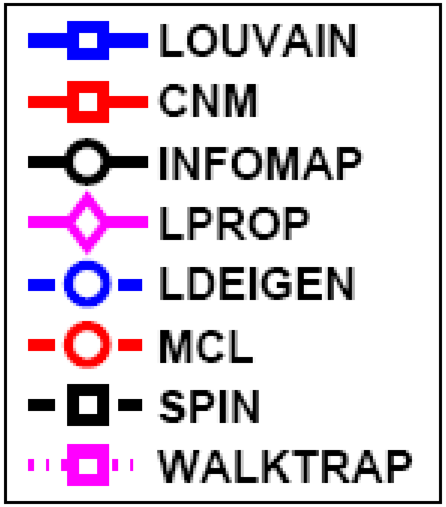
M. Rosvall, C. T. Bergstrom. Maps of information flow reveal community structure in complex networks. PNAS 105, 1118, 2008

WALKTRAP

Similarity computation between nodes using random walks.

P. Pons, M. Latapy. Computing communities in large networks using random walks. arXiv:physics/0512106, 2005

performance evaluation - empirical *2*



LPROP

Label propagation process.

U. N. Raghavan, R. Albert, S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76, 036106 (2007)

LDEIGEN

Use spectrum of modularity matrix.

M. Newman. Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 036104 (2006)

MCL

Markov Cluster algorithm: Use of "special" operations on adjacency matrix.

S. van Dongen. A cluster algorithm for graphs. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, May 2000

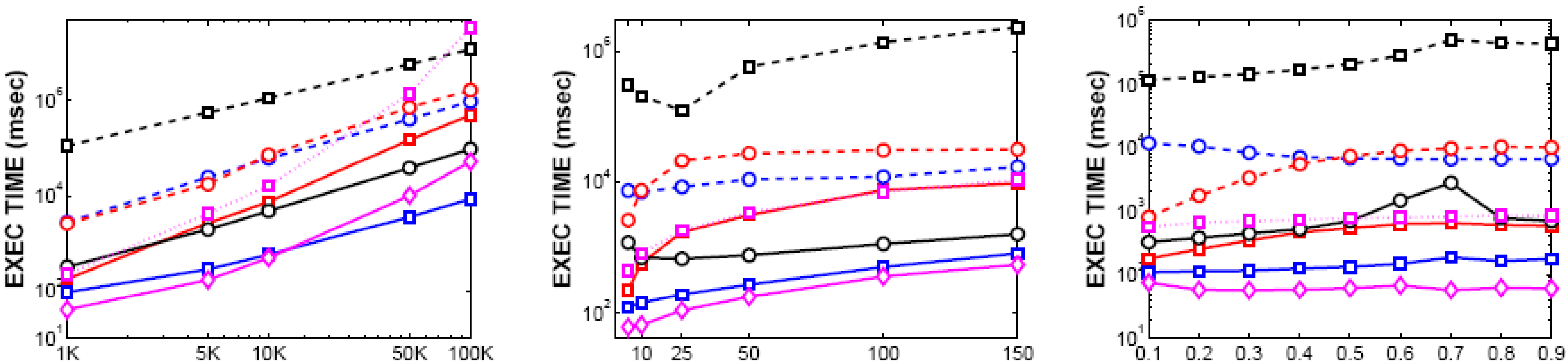
SPIN

Use models from statistical mechanics on the graph.

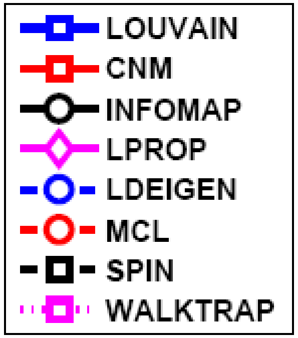
J. Reichardt, S. Bornholdt. Statistical mechanics of community detection. Phys. Rev. E 74, 016110, 2006

performance evaluation - empirical *3*

execution time

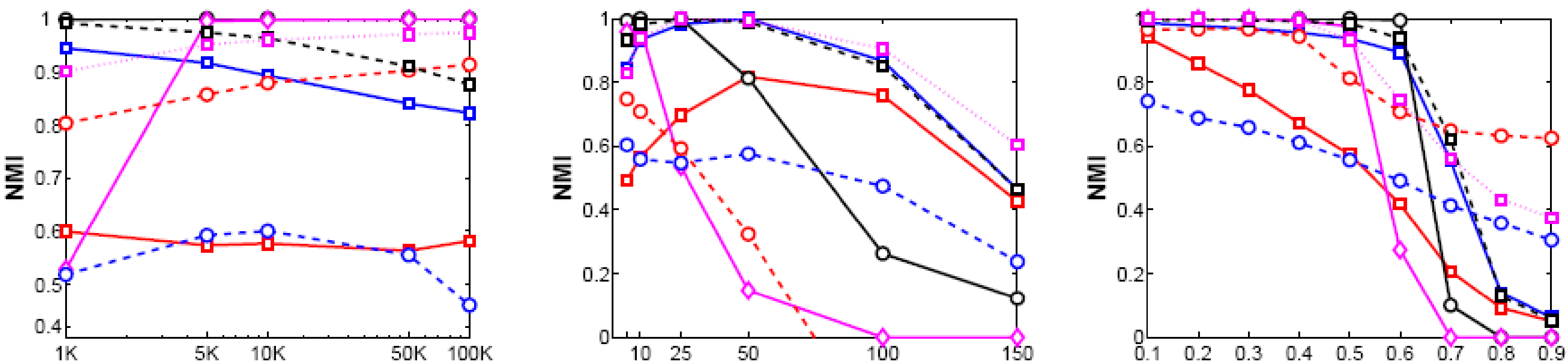


graph size | average node degree | mixing parameter

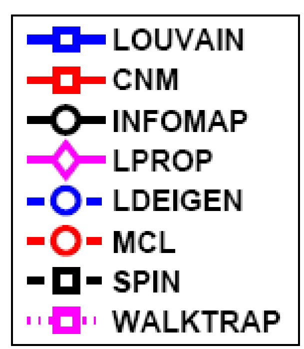


performance evaluation - empirical *4*

community structure precision (NMI)

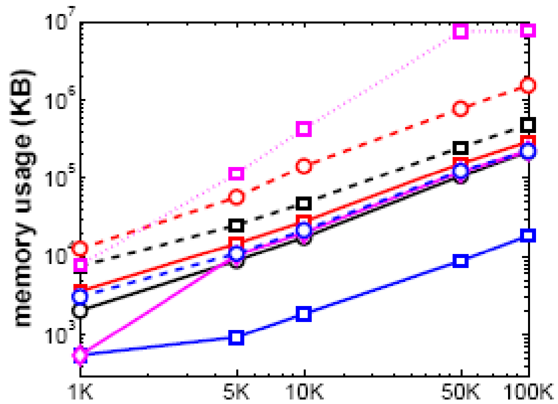


graph size | average node degree | mixing parameter

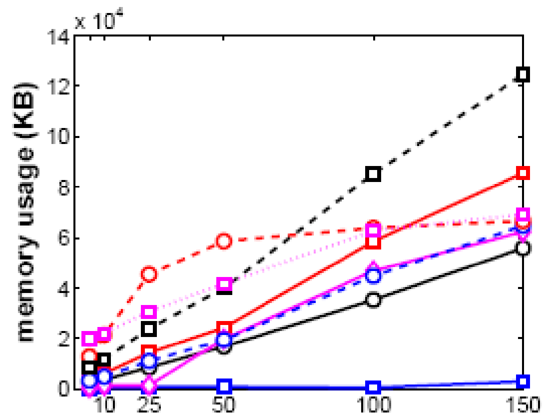


performance evaluation - empirical *5*

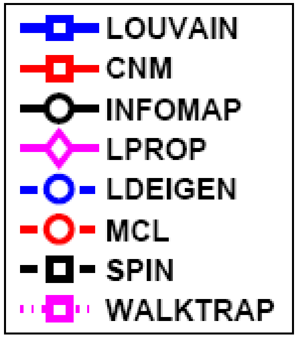
memory consumption



graph size



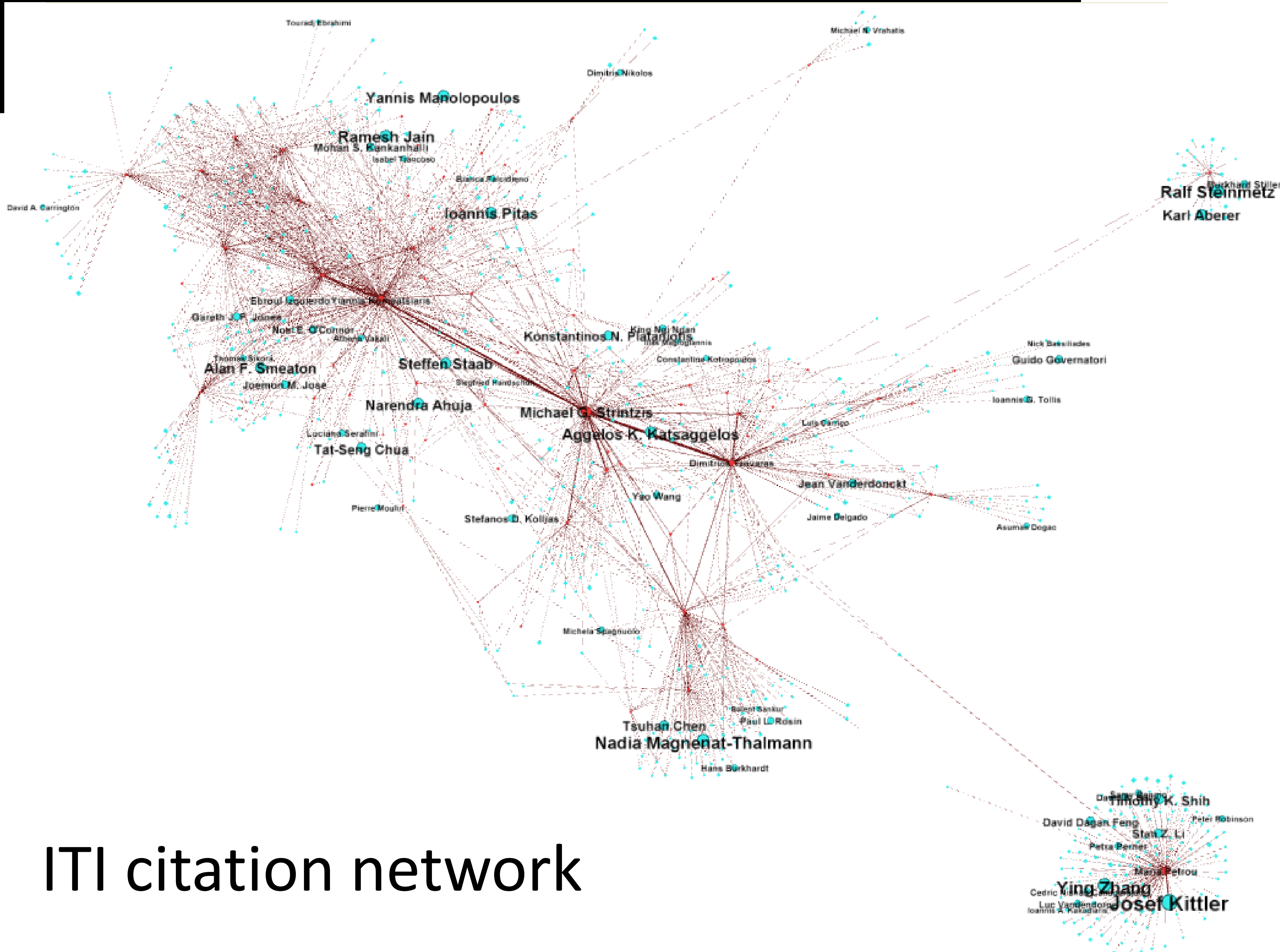
average node degree



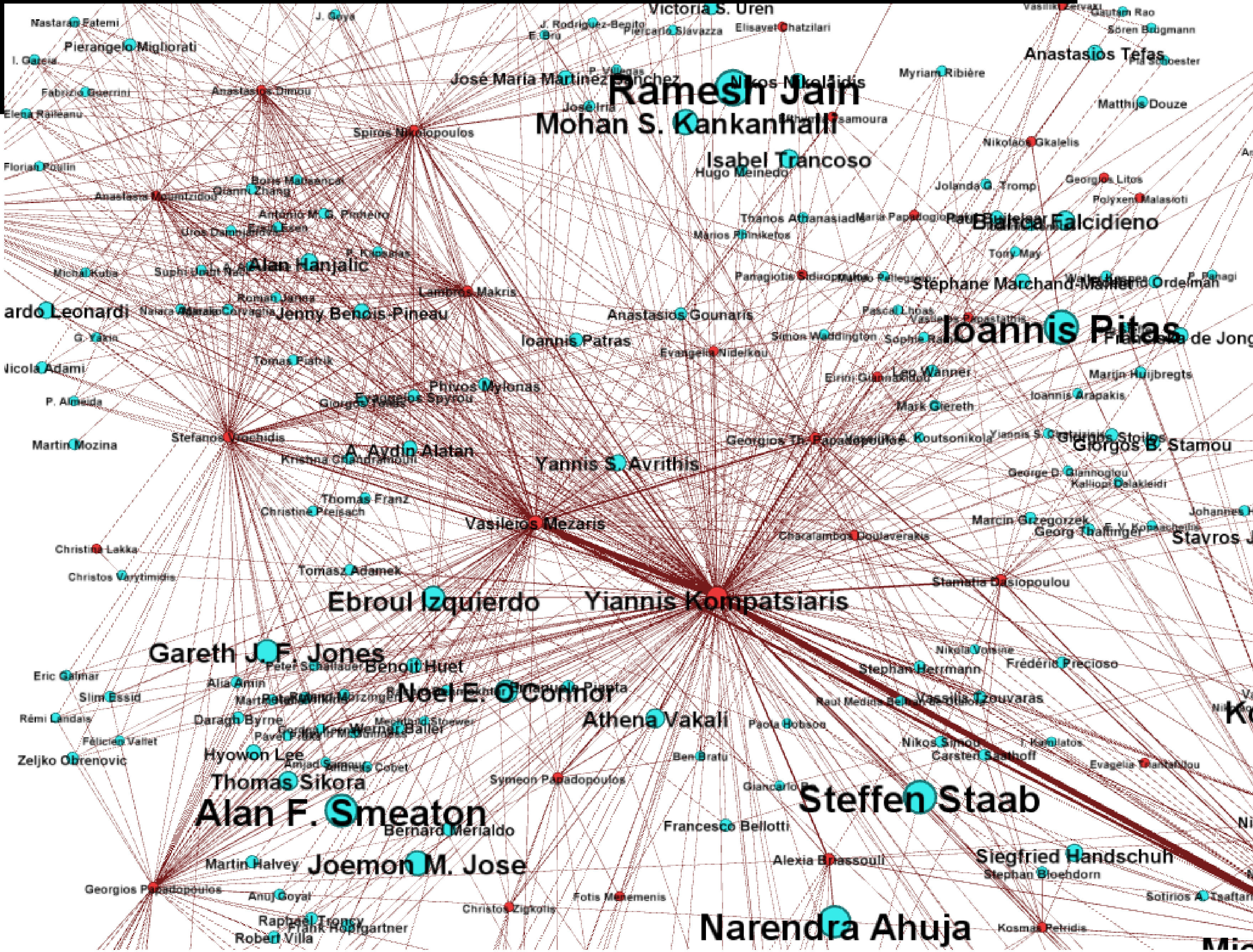
Citation networks, Social Tagging, Photo Clustering and POI Recommendation

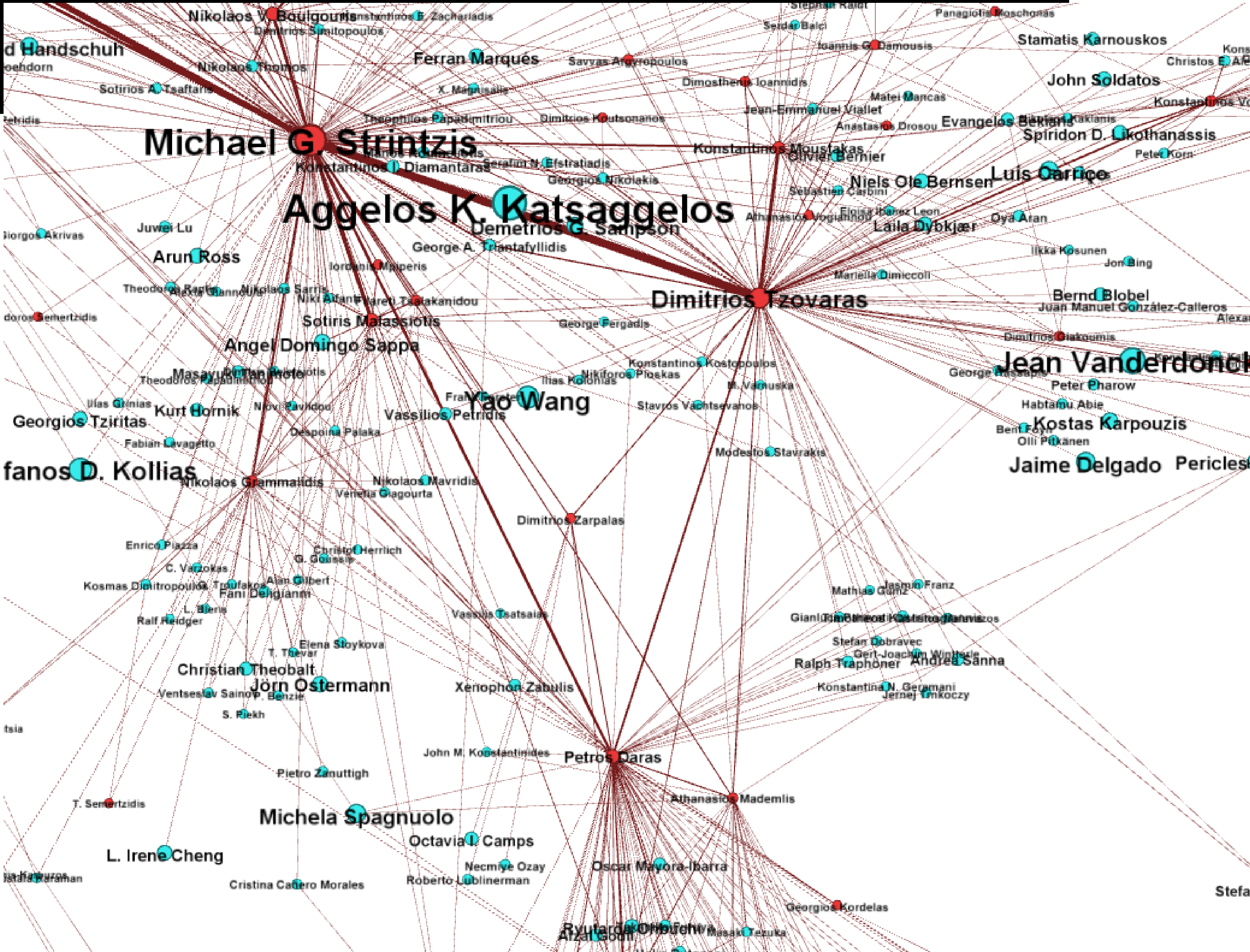
APPLICATIONS

#3

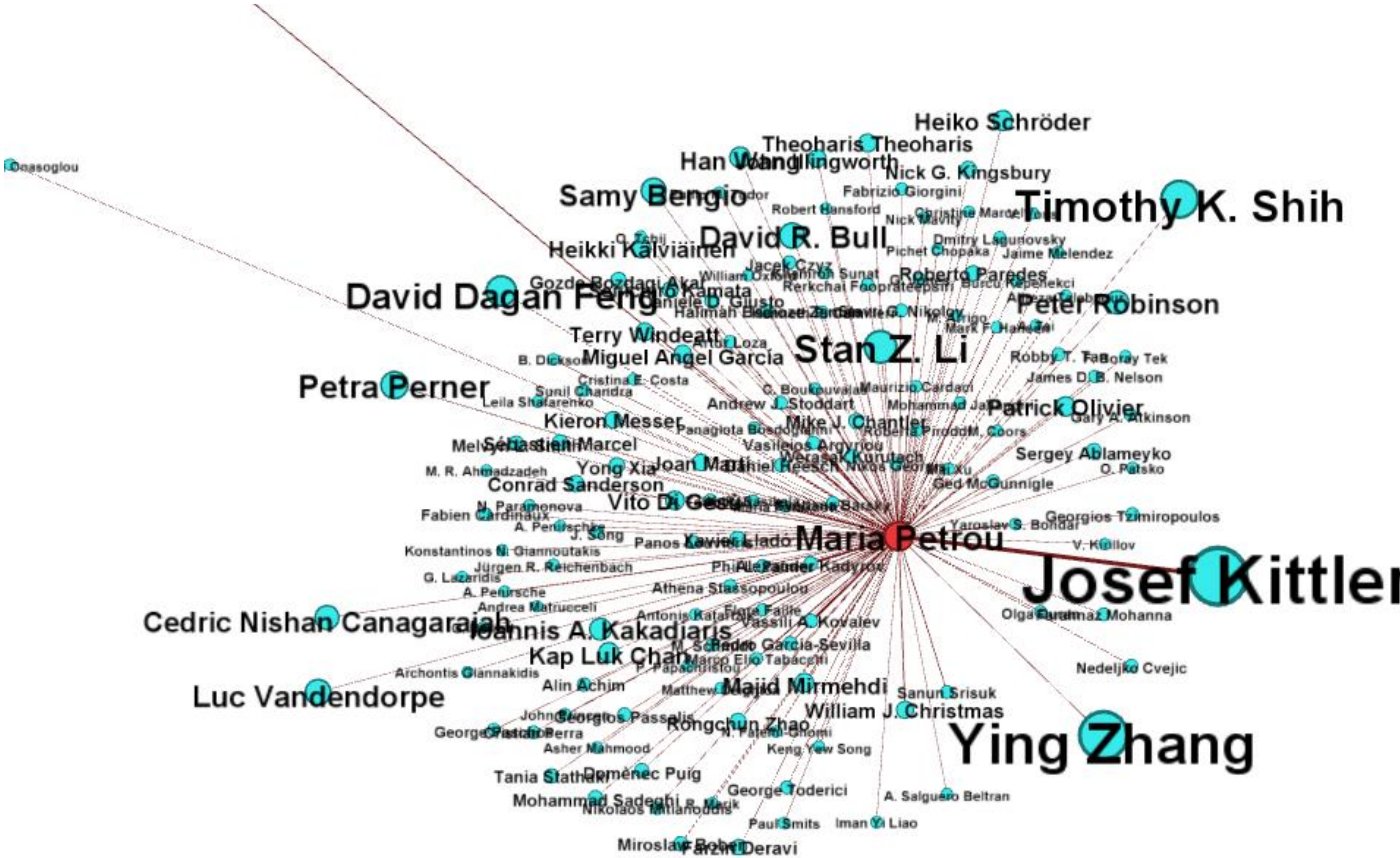


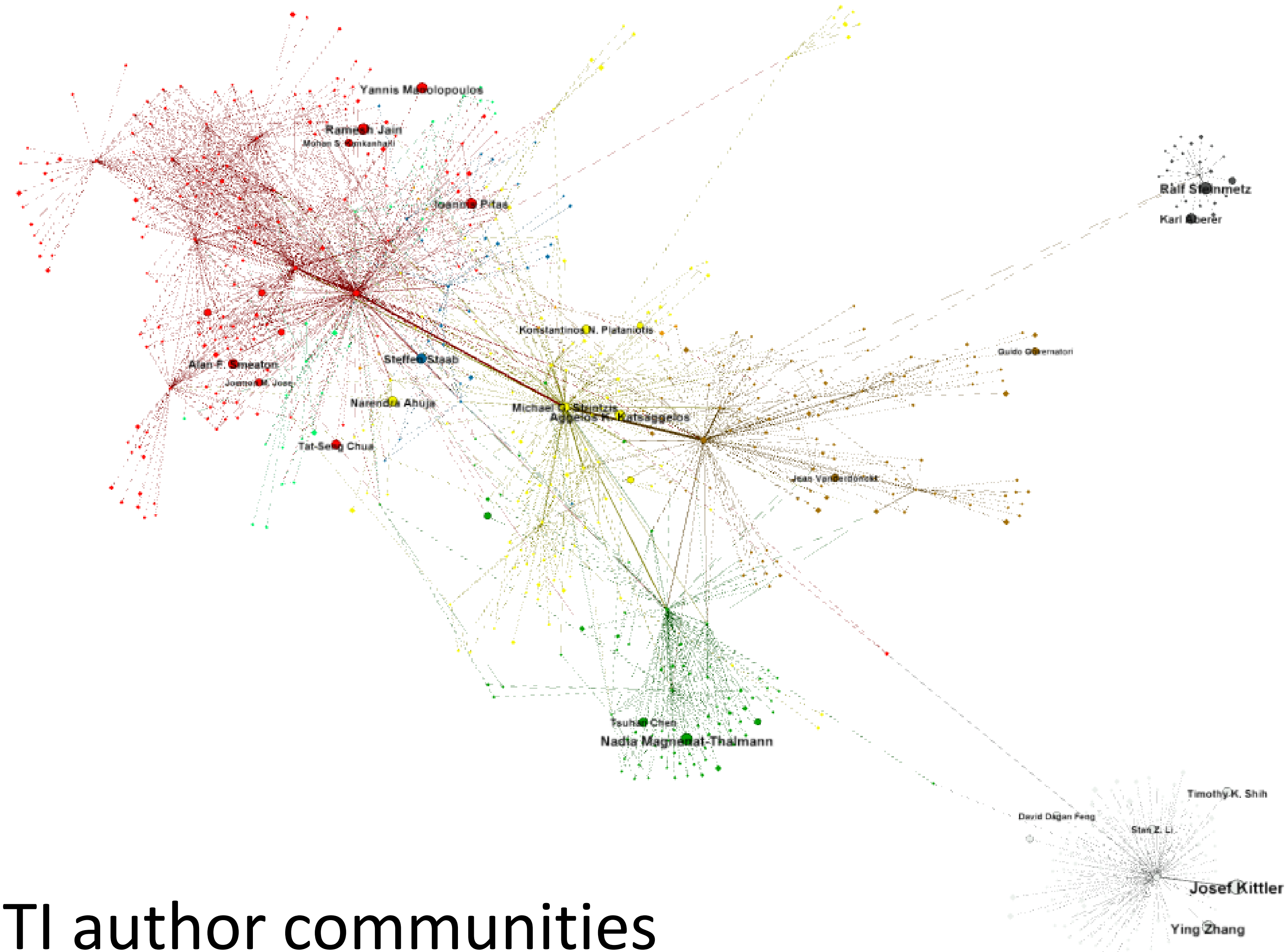
ITI citation network


Ramesh Jain
Mohan S. Kankanhalli
Isabel Trancoso
Victoria S. Uren
Anastasios Tefas
Bianca Falcidieno
Stephane Marchand-Maillet
Ioannis Pitas
Alan Hanjalic
Jenny Benois-Pineau
Anastasios Gounaris
Ioannis Patras
Phivos Mylonas
A. Aydin Alatan
Yannis S. Avrithis
Giorgos B. Stamou
Vasileios Mezaris
Yiannis Kompatsiaris
Ebroul Izquierdo
Gareth J. F. Jones
Benoit Huet
Noel E. O'Connor
Athena Vakali
Hyowon Lee
Thomas Sikora
Alan F. Smeaton
Bernard Merialdo
Joemon M. Jose
Steffen Staab
Siegfried Handschuh
Narendra Ahuja
Francesco Bellotti
Alexia Briassouli
Symeon Papadopoulos
Georgios Papadopoulos
Stefanos Vrochidis
Spiros Nikolopoulos
Anastasios Dimou
Lambros Makris
Stamatia Dasiopoulou
Georgios Th. Papadopoulos
Charalambos Doulaverakis
Stephan Herrmann
Frédéric Precioso
Thomas Franz
Tomasz Adamek
Martin Halvey
Anuj Goyal
Robert Villa
Christos Zigkolis
Fotis Menemenis
Kosmas Petridis
Zeljko Obrenovic
Martin Mozina
Nicola Adami
Pierangelo Migliorati
Matthijs Douze
Myriam Ribière
Marcin Grzegorzek
Georg Thallinger
Leo Wanner
Mark Greenwood
Marijn Huijbregts
Ioannis Arapakis
Ben Bratu
Paola Hobson
Stephan Bloehdorn
Sotirios A. Tsaftaris
Daragh Byrne
Alia Amin
Slim Essid
Rémi Landais
Eric Galmar
Félicien Vallet
Christina Lakka
Christos Varytimidis
Christine Preisach
Tomas Piatrik
Nikos Nikolaidis
José María Martínez Sánchez
Evangelia Nidelkou
Simon Waddington
Jolanda G. Tromp
Hugo Meinedo
Marios Phinikettos
Thanos Athanasiadis
Panagiotis Sidiropoulos
Nikos Simou
Carsten Saathoff
Evangelia Triantafillou
Anastasia Moumtzidou
Ioannis Arapakis
George D. Giannoglou


Michael G. Strintzis
Aggelos K. Katsaggelos
Dimitrios Tzovaras
Petros Daras
Jean Vanderdonck
Yao Wang
Nikolaos V. Boulgouris
Ferran Marqués
Stamatis Karnouskos
John Soldatos
Evangelos Bekiaris
Spiridon D. Likothanassis
Konstantinos Moustakas
Olivier Bernier
Niels Ole Bernsen
Luis Carriço
Laila Dybkjær
Konstantinos I. Diamantaras
Demetrios G. Sampson
George A. Triantafyllidis
Arun Ross
Juwei Lu
Sotiris Malassiotis
Angel Domingo Sappa
Georgios Tziritas
Kurt Hornik
Vassilios Petridis
Nikolaos Grammalidis
Dimitrios Zarpalas
Bernd Blobel
Juan Manuel González-Calleros
Dimitrios Giakoumis
Peter Pharow
Kostas Karpouzis
Jaime Delgado
Christian Theobalt
Jörn Ostermann
Xenophon Zabulis
Fani Deligianni
Ralph Traphöner
Andrea Sanna
Athanasios Mademlis
Michela Spagnuolo
Octavia I. Camps
L. Irene Cheng
Oscar Mayora-Ibarra
Georgios Kordelas
Cristina Cañero Morales
Roberto Lublinerman
Pietro Zanuttigh
Helin Dutagaci

ITI citation network





ITI author communities

ITI author communities

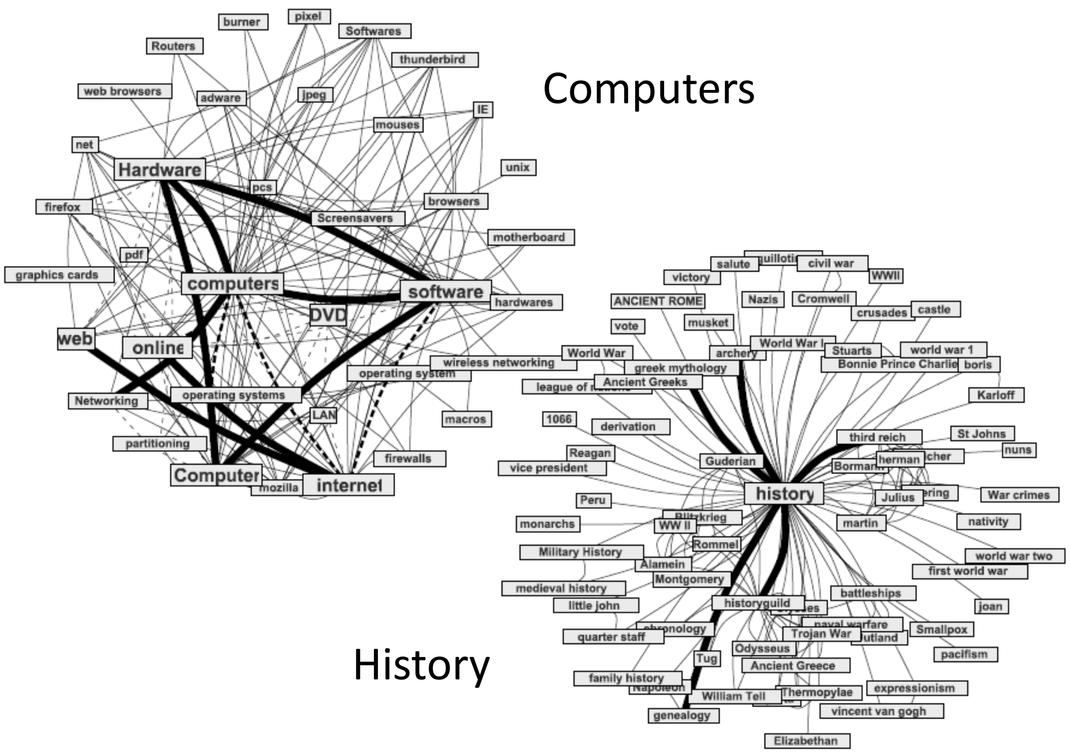
- Petrou
- Strintzis - Grammalidis - Malassiotis
- Tzovaras - Moustakas (not incl. some team members)
  - > Darlagiannis + ex-colleagues
- Kompatsiaris - Mezaris (not incl. some team members)
  - > Dasiopoulou + semantic web community (Staab et al.)
  - > S.Papadopoulos, Zigkolis + AUTH community (Vakali et al.)
  - > Tsampoulatidis
- Daras

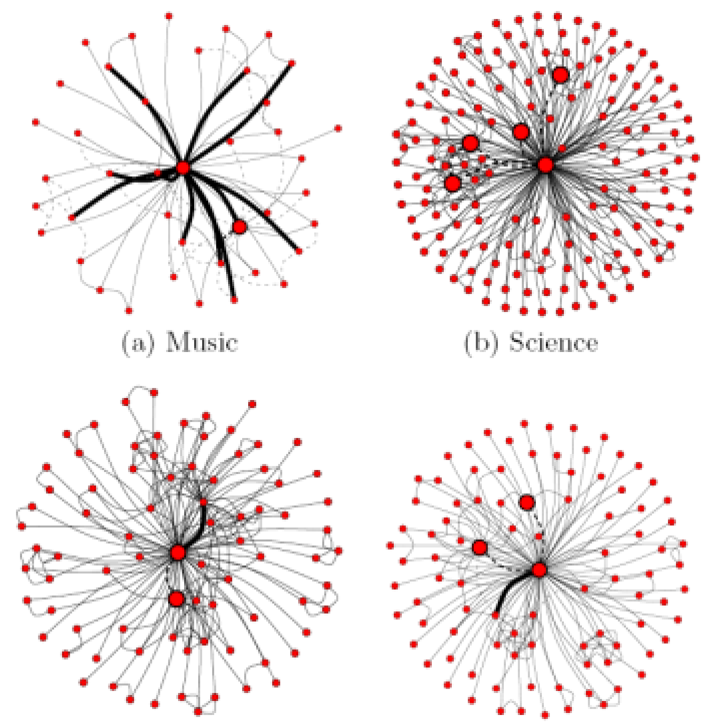
tag clustering

- In a social tagging system (delicious, flickr, BibSonomy), users annotate resources (articles, photos, citations) with **tags**.
- Using tag co-occurrence in the context of resources, we can create **tag graphs**.
- We can then apply community detection methods on such graphs.
- The extracted communities usually correspond to topics that are of interest to the tagging system under study.

tag clustering - example 1

Topic of interest detection by means of local community detection on tag graphs (Lycos IQ, questions & answers).





(a) Music (b) Science

(c) Film (d) Animals

S. Papadopoulos, A. Skusa, A. Vakali, Y. Kompatsiaris, N. Wagner. “Bridge Bounding: A Local Approach for Efficient Community Discovery in Complex Networks.” In arXiv:0902.0871.

tag clustering - example 2

(a) Basic folksonomy statistics

Dataset	#triplets	U	R	T
BIBSONOMY-200K	234,403	1,185	64,119	12,216
FLICKR-1M	927,473	5,463	123,585	27,969
DELICIOUS-7M	7,501,032	112,950	1,332,796	251,352

(b) Tag graph statistics (for large component)

Dataset	$\|V\|$	$\|E\|$	$\bar{k}$	$\overline{cc}$
BIBSONOMY-200K	11,949	236,791	39.63	0.6689
FLICKR-1M	27,521	693,412	50.39	0.8512
DELICIOUS-7M	216,844	3,443,367	31.76	0.8018

					R-1M		DELICIOUS-7M		
					N	HGC	CNM	SCAN	HGC
					5			56,893	
					3	49,851	56,166	13,974	33,107
					9	**11,454**	1,022	3,624	**6,258**
U_{TP}	189	717	**837**	305	1,399	**1,666**	459	1,506	**2,628**
P (%)	1.81	**22.59**	11.90	3.73	**46.38**	22.98	1.82	**25,93**	18.90
R (%)	1.79	7.36	**9.24**	3.71	18.65	**20.50**	1.80	6.37	**11.00**
F (%)	1.80	**11.10**	10.40	3.72	**26.60**	21.67	1.81	10.23	**13.91**
$P@1$ (%)	1.68	3.96	**5.09**	1.95	8.02	**9.85**	1.64	2.78	**7.95**
$P@5$ (%)	2.18	**29.06**	17.27	3.41	**46.84**	21.27	2.35	**36.91**	29.49

S. Papadopoulos, Y. Kompatsiaris, A. Vakali. "Leveraging Collective Intelligence through Community Detection in Tag Networks." In Proceedings of CKCaR'09 Workshop on Collective Knowledge Capturing and Representation, Redondo Beach, California, USA (September 1, 2009)

S. Papadopoulos, Y. Kompatsiaris, A. Vakali. "A Graph-based Clustering Scheme for Identifying Related Tags in Folksonomies." In Proceedings of DaWaK'10, 12th International Conference on Data Warehousing and Knowledge discovery (Bilbao, Spain), Springer-Verlag, 65-76

tag disambiguation

tag co-occurrence graph

↓

overlapping community detection

↓

senses

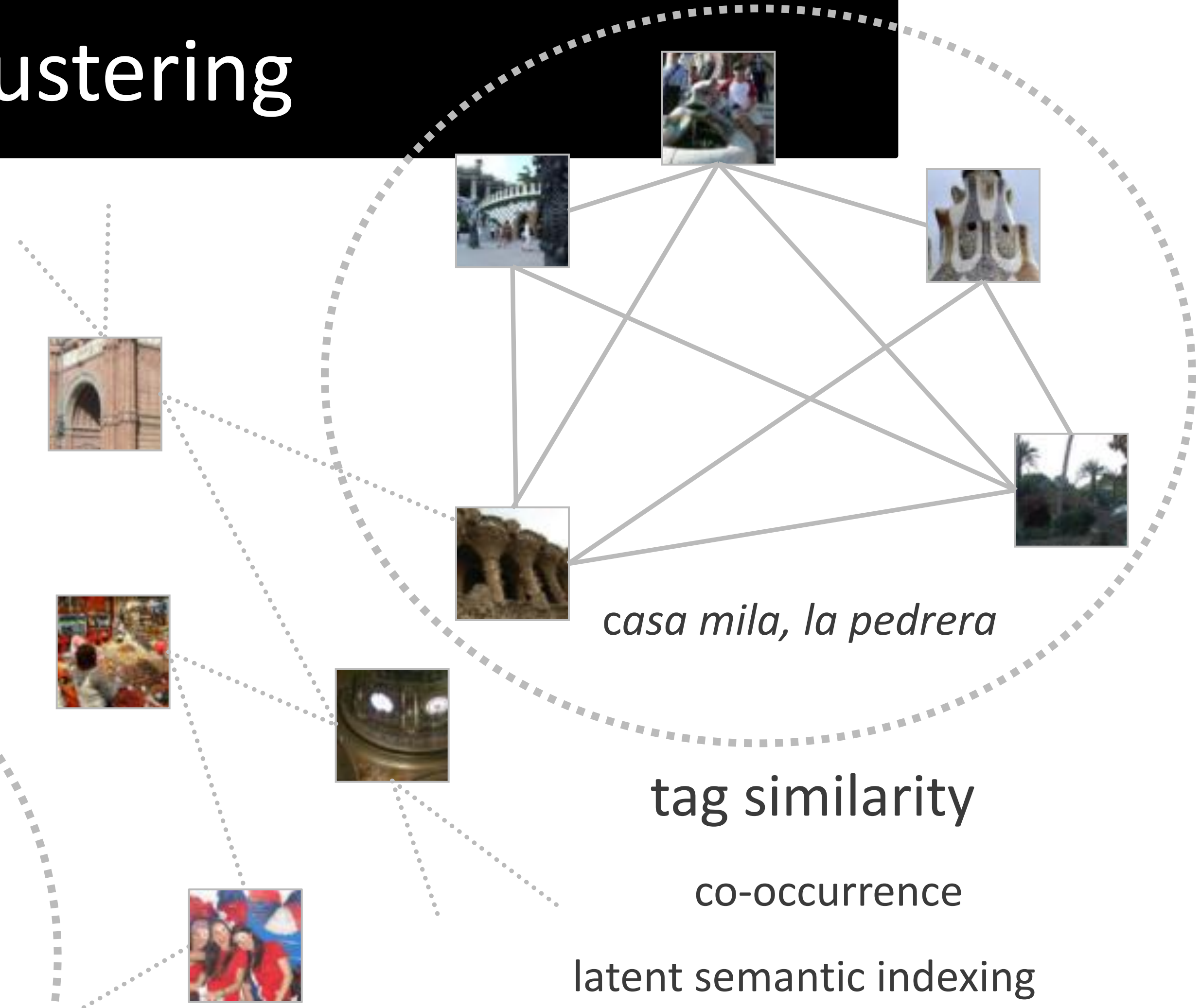
S. Papadopoulos, A. Vakali, Y. Kompatsiaris. "Community Detection in Collaborative Tagging Systems." In Book Community-built Database: Research and Development, Springer (2011)

	networks (BIBSONOMY-200K)
Network analysis	socialanalysis, socialdynamics, sociophysics, autocorrelation, swarm, percolation, selforg, fitnesslandscapes, socialstructure, ...
Epidemic spread	large, spreading, enidtl, kcore, dtl, pysics, epidemic, gda, vespignani, networks, delivery, pastorsatorras
Bibliometrics	citation, reputation, publication, bibliometrics, academics, impact, ...
VoIP networks	voip, phone, skype, telephony, jajah, vonage, internettelephony, ...
Security	proxy, encryption, wlan, password, openssl, cracker, https, wep, ...
	bridge (FLCIKR-1M)
Bac de Roda	bacderoda, puente de calatrava, bru, puente luminoso, structural, ...
Bridge of Sighs (Bisbe)	carrer del bisbe, bridge of sighs, footbridge, cobblestone, bascule, ...
Port bridge	port, puerto, sea, port vell, ship, boat
Pont del Centenari	transporte, pont, montserrat, new
	bean (DELICIOUS-7M)
Plant seeds	lentils, blackbean, greenbean, kidneybeans, hummus, chickpeas, ...
Coffee beans	grown, arabica, robusta, pricey, coffea, wholesaler
Bean bag	toffet, velveteen, crushed, swirl, tuffet, herbagreeen, swirled
Java beans	spring, hibernate, jsf, ejb
Travis Bean	guitars, fender, musicstuff, giger, telecaster, strat, stratocaster, ...
Frances Bean Cobain	courtneylove, cobain, francesbean, lindacarroll, thevoid, girlgerms, ...

Flickr photo clustering

SURF

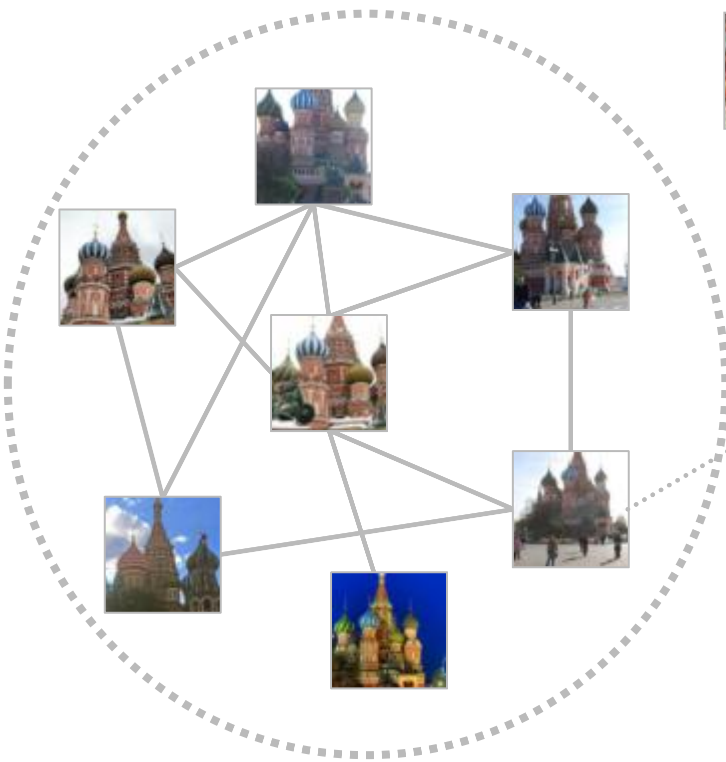
visual similarity



tag similarity

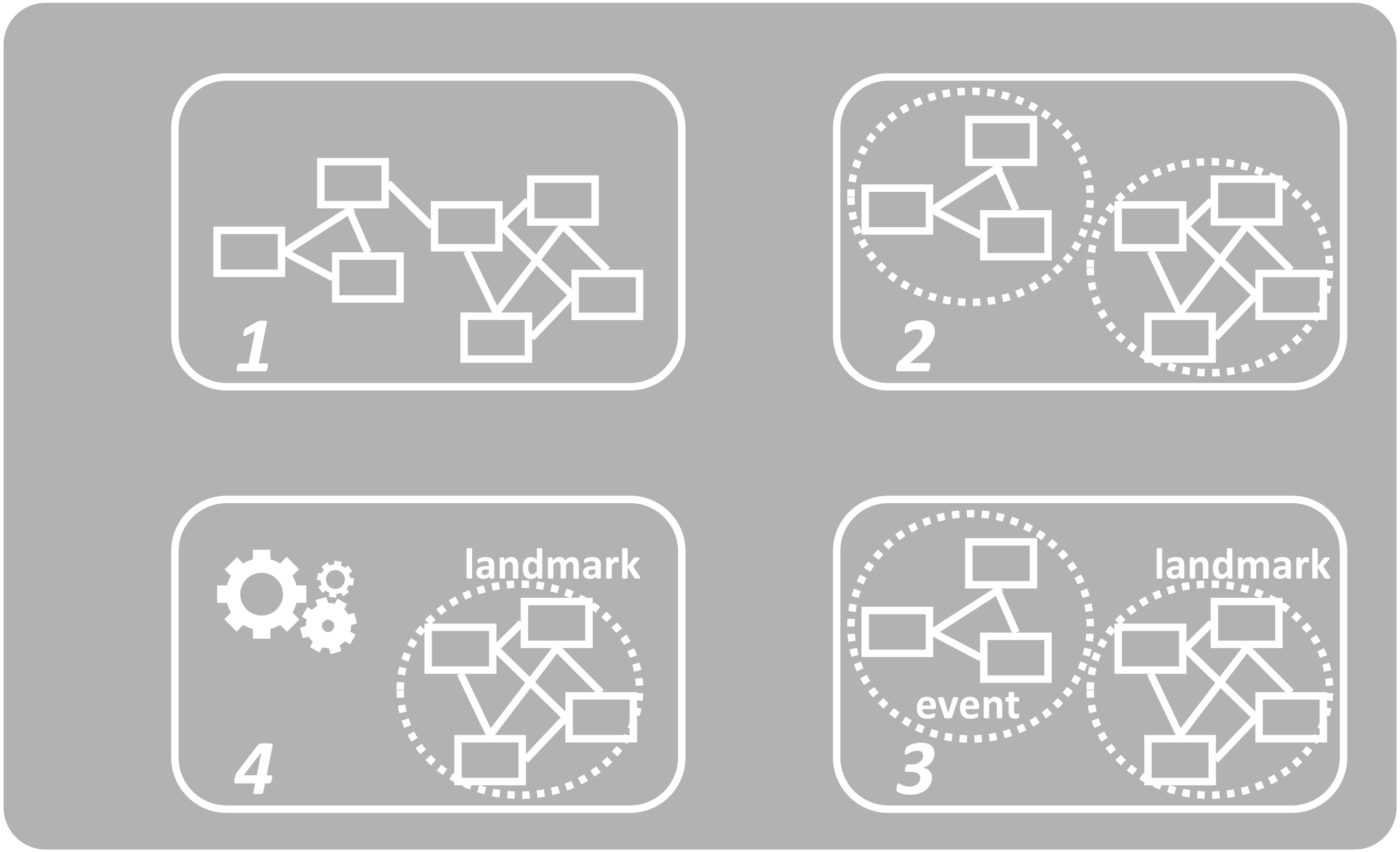
co-occurrence

latent semantic indexing



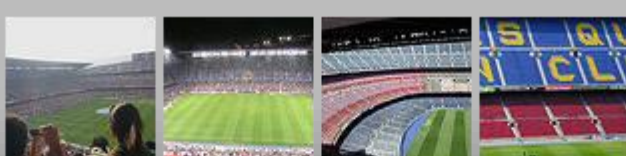
S. Papadopoulos, C. Zigkolis, G. Tolias, Y. Kalantidis, P. Mylonas, Y. Kompatsiaris, A. Vakali. “Image Clustering through Community Detection on Hybrid Image Similarity Graphs.” In Proceedings of ICIP 2010, International Conference on Image Processing (Hong Kong), IEEE, pp. 2353-2356

landmark & event detection *1*



S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, A. Vakali. "Cluster-based Landmark and Event Detection on Tagged Photo Collections." In IEEE Multimedia Magazine 18(1), pp. 52-63, 2011

landmark & event detection 2

	sagrada familia, cathedral, catholic	15.2m
	la pedrera, casa mila	31.8m
	parc guell	9.6m
	boqueria, market, mercado, ramblas	82.1m
	camp nou, fc barcelona, nou camp	18.7m

landmark & event detection *3*

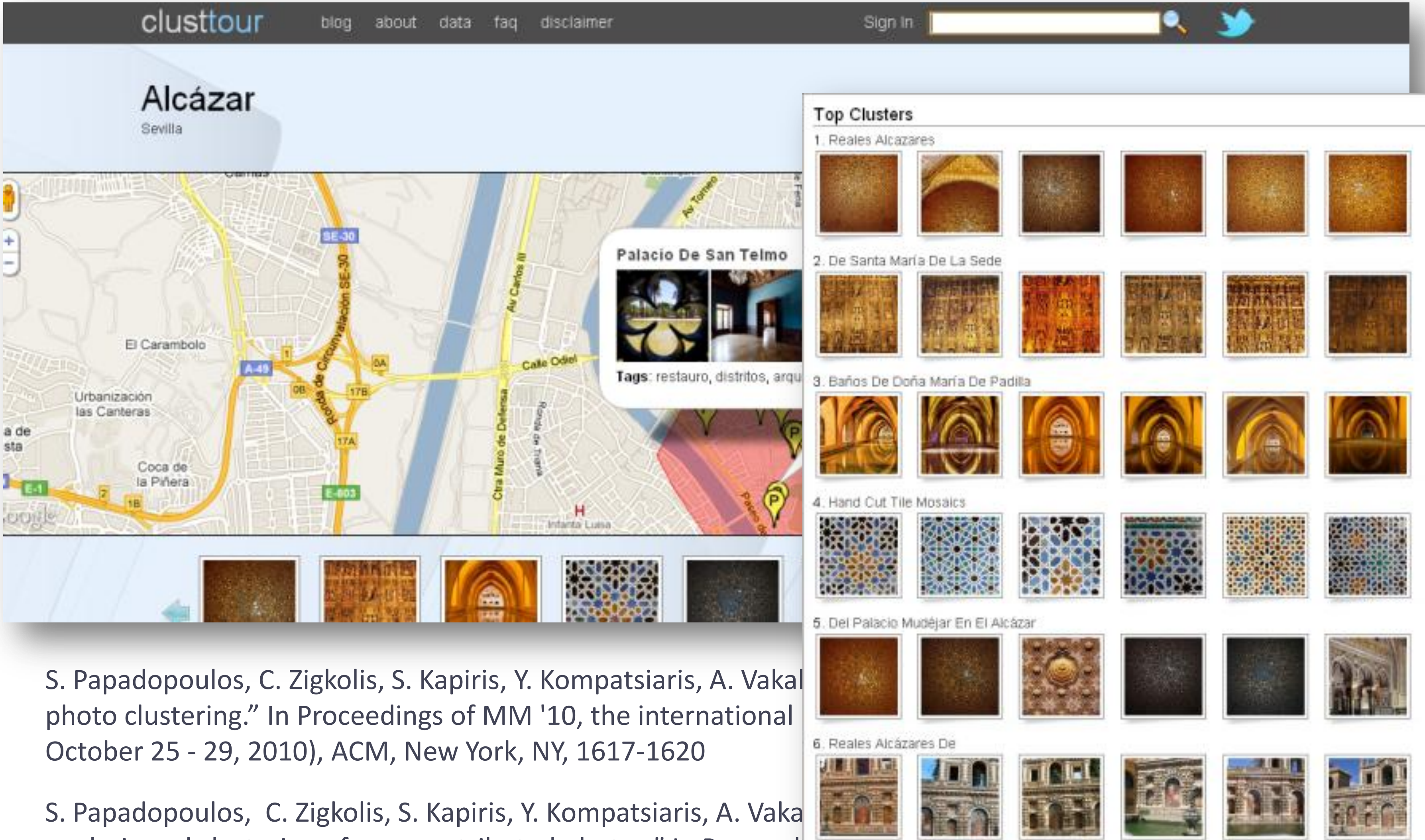
music, concert, gigs, dj 43.1%

conference, presentation 6.5%

local traditional, parades 4.6%

racing, motorbikes, f1 3.3%

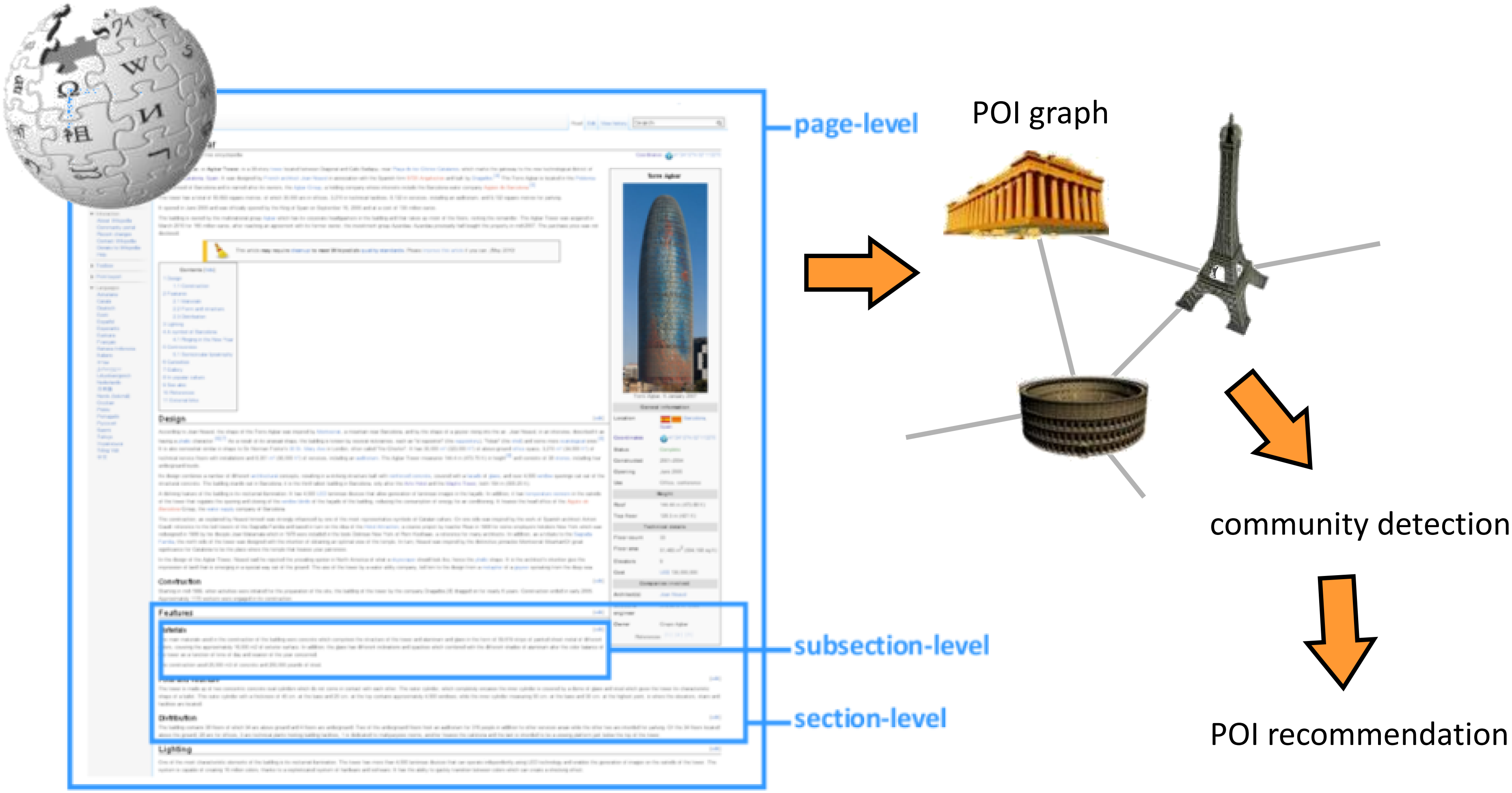
cluster-based city exploration



S. Papadopoulos, C. Zigkolis, S. Kapiris, Y. Kompatsiaris, A. Vakal photo clustering." In Proceedings of MM '10, the international October 25 - 29, 2010), ACM, New York, NY, 1617-1620

S. Papadopoulos, C. Zigkolis, S. Kapiris, Y. Kompatsiaris, A. Vaka analysis and clustering of user contributed photos." In Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR '11). ACM, New York, NY, USA, , Article 65 , 2 pages.

POI recommendation



M. Janik, S. Papadopoulos, B. Sigurbjornsson. “D3.3 Mass Classification and Clustering.” Technical report, WeKnowIt project, 2010.

conclusions

- Network:

  an omni-present model of relational structure
- Community detection:

  a valuable tool for understanding structure in massive networks
- Applications:
  > Classic social networks analysis (e.g. citation graph)

  > Multimedia clustering and mining
- Challenges
  > scalability, dynamic networks

  > interpretation of results, evaluation

questions