ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό

Ενότητα # 5: Community Detection and Evaluation in Social and Information Networks

Διδάσκων: Μιχάλης Βαζιργιάννης

Τμήμα: Προπτυχιακό Πρόγραμμα Σπουδών "Πληροφορικής"

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Οικονομικό Πανεπιστήμιο Αθηνών**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ
2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Οι εικόνες προέρχονται … .

Σκοποί ενότητας

Εισαγωγή και εξοικείωση με τις μεθόδους Introduction & Motivation, Community evaluation measures, Graph clustering algorithms, Alternative Methods for Community Evaluation, New directions for research in the area of graph mining.

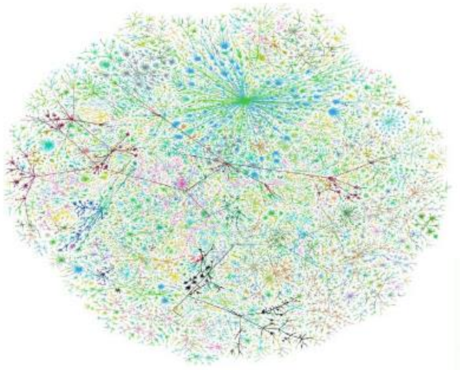
Περιεχόμενα ενότητας

- Introduction & Motivation
- Community evaluation measures
- Graph clustering algorithms
- Alternative Methods for Community Evaluation
- New directions for research in the area of graph mining

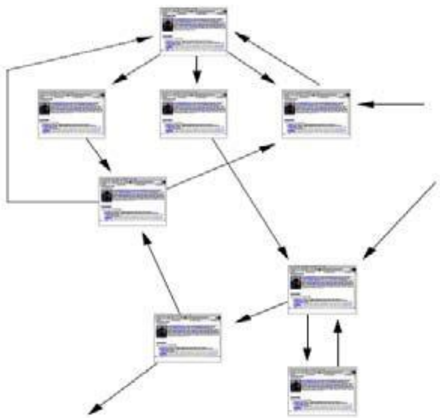
Introduction & Motivation

Μάθημα: Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό
Ενότητα # 5: Community Detection and Evaluation
in Social and Information Networks
Διδάσκων: Μιχάλης Βαζιργιάννης
Τμήμα: Προπτυχιακό Πρόγραμμα Σπουδών "Πληροφορικής"

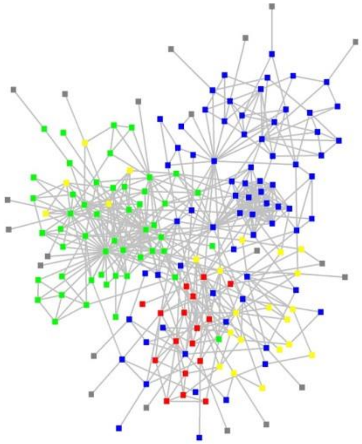
Networks are Everywhere



(a) Internet

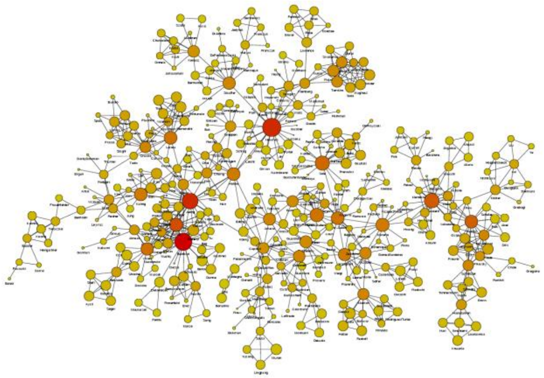


(b) World Wide Web

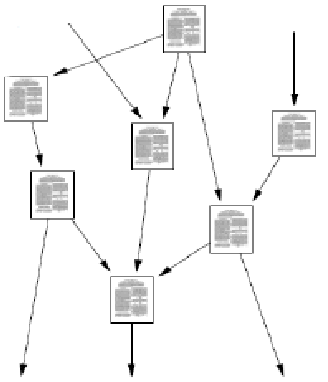


(c) Email network

(d) Social network



(e) Collaboration network



(f) Citation network

Social Networks Growth

- Social networking accounts for 1 of every 6 minutes spent online [http://blog.comscore.com/]
- One out of seven people on Earth is on Facebook
- People on Facebook install 20 million "Apps" every day
- YouTube has more than on billion unique users who visit every month (Oct. 2014)
- Users on YouTube spend a total of 6 billion hours per month (almost an hour for every person on Earth!
- Wikipedia hosts ~34 million articles and has over 91,000 contributors
- 500 million average Tweets per day occur on Twitter (Oct. 2014)

[http://www.jeffbullas.com/2011/09/02/20-stunning-social-media-tatistics/#q3eTJhr64rtD0tLF.99]

Communities in Real Networks

- Real networks are not **random graphs** (e.g., the Erdos-Renyi random graph model)
- Present fascinating patterns and properties:
  - The **degree distribution** is skewed, following a power-law
  - The **average distance** between the nodes of the network is short (the small-world phenomenon)
  - The edges between the nodes may not represent reciprocal relations, forming **directed networks with non-symmetric links**
  - **Edge density** is inhomogeneous (groups of nodes with high concentration of edges within them and low concentration between different groups

Community Detection

- **Community detection** in graphs aims to identify the modules and, possibly, their hierarchical organization, using mainly the information encoded in the graph topology
- First attempt dates back to 1955 by Weiss and Jacobson searching for work groups within a government agency

Communities – application domains

- **Social communities** have been studied for a long time (Coleman, 1964; Freeman, 2004; Kottak, 2004; Moody and White, 2003)
- **In biology**: protein-protein interaction networks, communities are likely to group proteins having the same specific function within the cell (Chen, 2006; Rives and Galitski 2003; Spirin and Mirny, 2003)
- **World Wide Web**: communities correspond to groups of pages dealing with the same or related topics (Dourisboure et al., 2007; Flake et al., 2002)
- **Metabolic networks** they may be related to functional modules such as cycles and pathways (Guimera and Amaral, 2005; Palla et al., 2005)
- In **food webs** they may identify compartments (Krause et al., 2003; Pimm, 1979)

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Community evaluation measures

Μάθημα: Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό

Ενότητα # 5: Community Detection and Evaluation in Social and Information Networks

Διδάσκων: Μιχάλης Βαζιργιάννης

Τμήμα: Προπτυχιακό Πρόγραμμα Σπουδών “Πληροφορικής”

Basics

- The notion of **community structure** captures the tendency of nodes to be organized into modules (communities, clusters, groups)
  - Members within a community are **more similar** among each other
- Typically, the communities in graphs (networks) correspond to **densely connected** entities (nodes)
- Set of nodes with **more/better/stronger** connections between its members, than to the rest of the network
- Why this happens?
  - Individuals are typically organized into social groups (e.g., family, associations, profession)
  - Web pages can form groups according to their topic
  - …

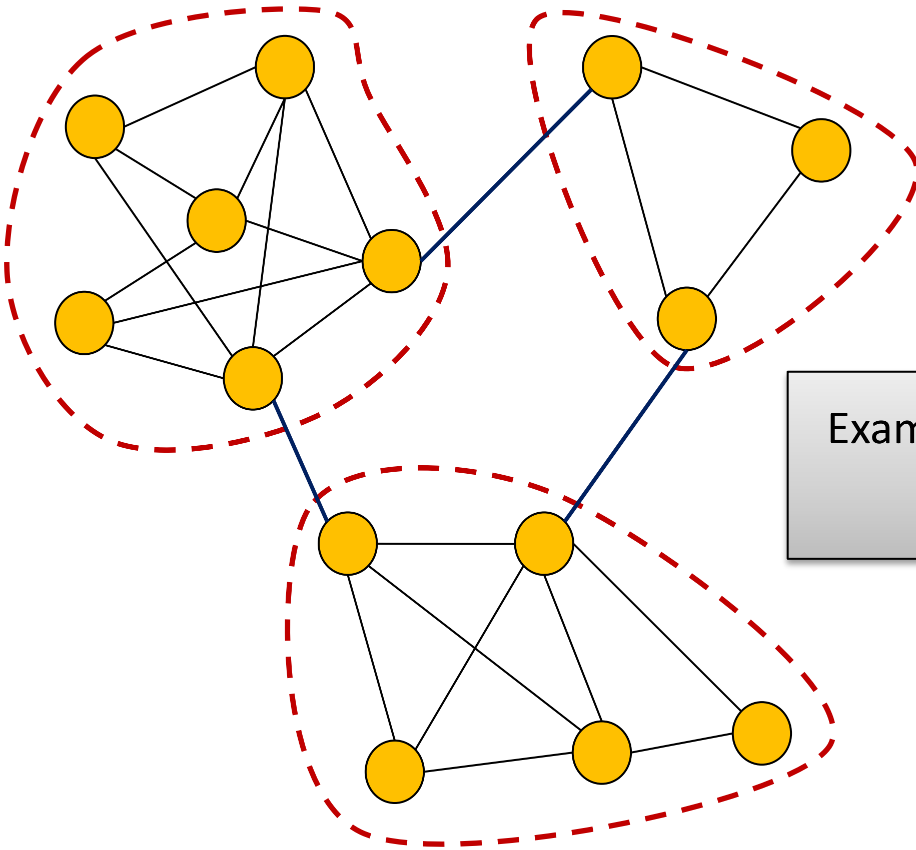
Definition/notion of communities

- How a community in graphs looks like?
- The property of community structure is **difficult** to be defined
  - There is no universal definition of the problem
  - It depends heavily on the application domain and the properties of the graph under consideration
- Most widely used notion/definition of communities is based on the number of edges within a group (density) compared to the number of edges between different groups

A community corresponds to a group of nodes with more **intra-cluster** edges than **inter-clusters** edges

[Newman '03], [Newman and Girvan '04], [Schaeffer '07], [Fortunato '10],
[Danon et al. '05], [Coscia et al. 11]

Schematic representation of communities



Example graph with three communities

Community detection in graphs

- How can we extract the inherent communities of graphs?
- Typically, a two-step approach
  1. Specify a **quality measure** (evaluation measure, objective function) that quantifies the desired properties of communities
  2. Apply **algorithmic techniques** to assign the nodes of graph into communities, optimizing the objective function
- Several measures for quantifying the quality of communities have been proposed
- They mostly consider that communities are set of nodes with many edges between them and few connections with nodes of different communities
  - Many possible ways to formalize it

Community evaluation measures

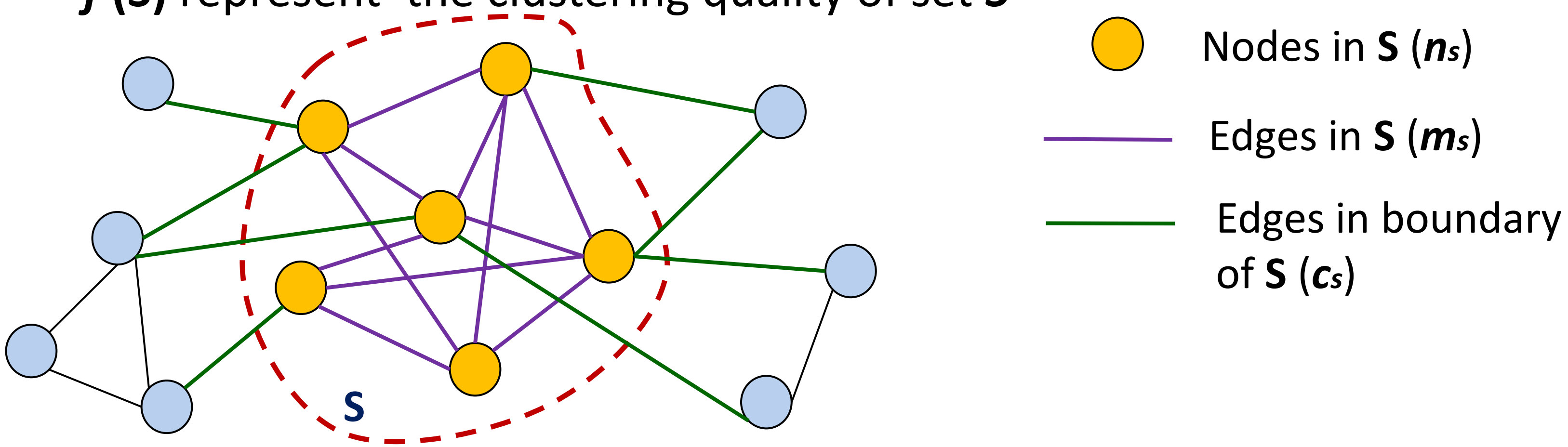
- Focus on
  - Intra-cluster edge density (# of edges within community),
  - Inter-cluster edge density (# of edges across communities)
  - Both two criteria

- We group the community evaluation measures according to
  - Evaluation based on **internal** connectivity
  - Evaluation based on **external** connectivity
  - Evaluation based on **internal and external** connectivity
  - Evaluation based on **network model**

[Leskovec et al. '10], [Yang and Leskovec '12], [Fortunato '10]

Notation

- **G = (V, E)** is an undirected graph, **|V| = n**, **|E| = m**
- **S** is the set of nodes in the cluster
- $\boldsymbol{n_s}$ **= |S|** is the number of nodes in **S**
- $\boldsymbol{m_s}$ is the number of edges in **S**, $m_s = \left| \{(u,v): u \in S, v \in S\} \right|$
- $\boldsymbol{c_s}$ is the number of edges on the boundary of **S**, $c_s = \left| \{(u,v): u \in S, v \notin S\} \right|$
- $\boldsymbol{d_u}$ is the degree of node $\boldsymbol{u}$
- ***f*** **(S)** represent the clustering quality of set ***S***

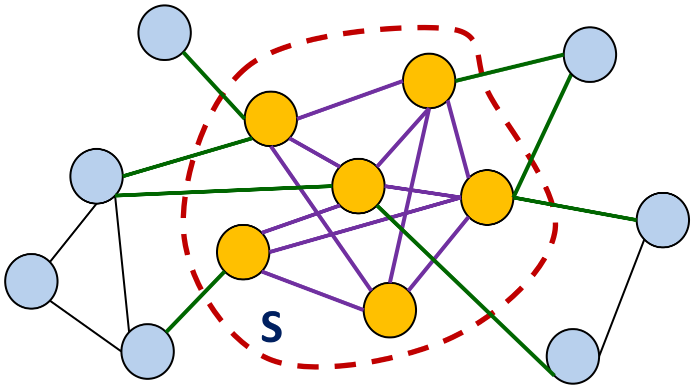


Evaluation based on internal connectivity (1)

- **Internal density [Radicchi et al. '04]**

$$f(S) = \frac{m_s}{n_s(n_s - 1)/2}$$

Captures the internal edge density of community **S**



- **Edges inside [Radicchi et al. '04]**

$$f(S) = m_s$$

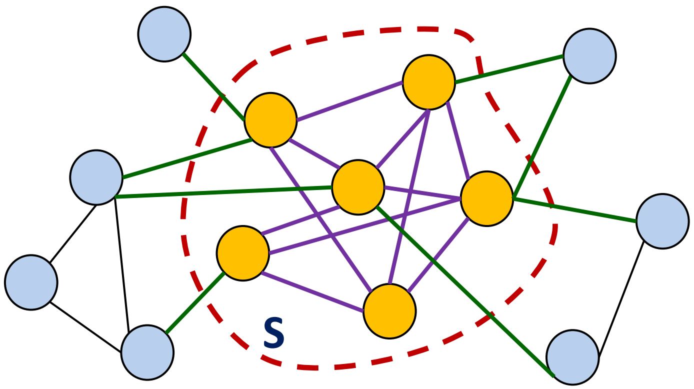
Number of edges between the nodes of **S**

Evaluation based on internal connectivity (2)

- **Average degree [Radicchi et al. '04]**

$$f(S) = \frac{2m_s}{n_s}$$

Average internal degree of nodes in **S**



- **Fraction over median degree (FOMD) [Yang and Leskovec '12]**

$$f(S) = \frac{\left|\left\{u : u \in S, \left|\{(u,v) : v \in S\}\right| > d_m\right\}\right|}{n_s}$$

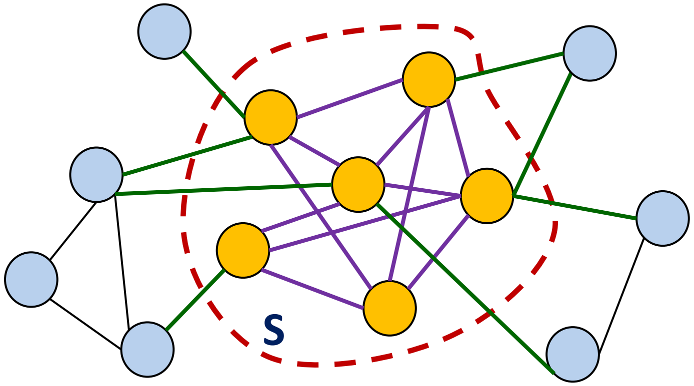
Fraction of nodes in **S** with internal degree greater than $\boldsymbol{d_m}$, where $\boldsymbol{d_m}$ **= median (**$\boldsymbol{d_u}$**)**

Evaluation based on internal connectivity (3)

- **Triangle participation ratio (TPR) [Yang and Leskovec ’12]**

$$f(S)=\frac{\left|\left\{u:u\in S,\left\{(v,w):v,w\in S,(u,v)\in E,(u,w)\in E,(v,w)\in E\right\}\neq\emptyset\right\}\right|}{n_s}$$

Fraction of nodes in **S** that belong to a triangle

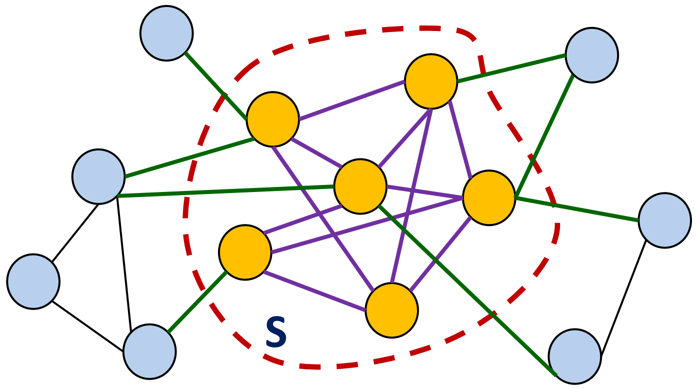


Evaluation based on external connectivity

- **Expansion [Radicchi et al. '04]**

$$f(S) = \frac{c_s}{n_s}$$

Measures the number of edges per node that point outside **S**



- **Cut ratio [Fortunato '10]**

$$f(S) = \frac{c_s}{n_s(n - n_s)}$$

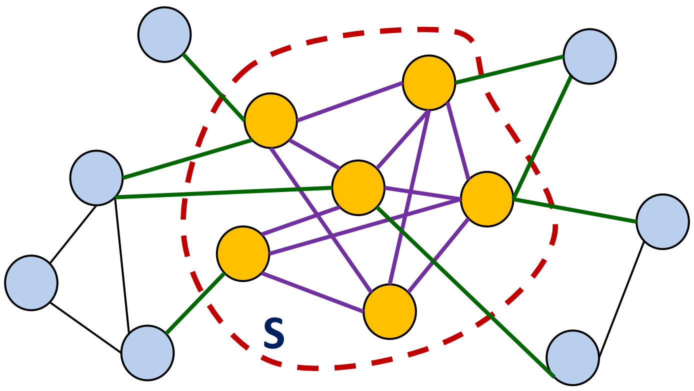
Fraction of existing edges – out of all possible edges – that leaving **S**

Evaluation based on internal and external connectivity (1)

- **Conductance [Chung '97]**

$$f(S) = \frac{c_s}{2m_s + c_s}$$

Measures the fraction of total edge volume that points outside **S**



- **Normalized cut [Shi and Malic '00]**

$$f(S) = \frac{c_s}{2m_s + c_s} + \frac{c_s}{2(m - m_s) + c_s}$$

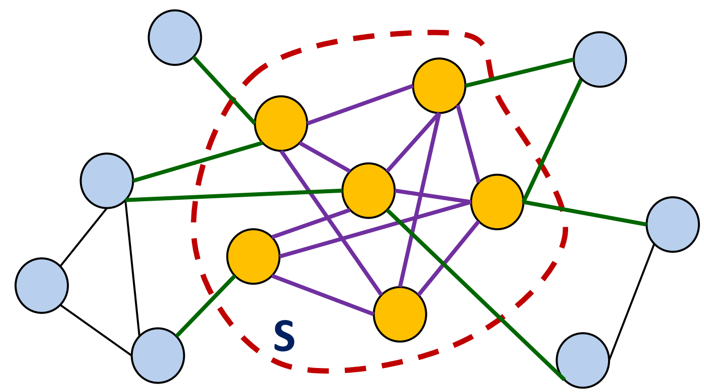
Measures the fraction of total edge volume that points outside **S** normalized by the size of **S**

Evaluation based on internal and external connectivity (2)

- **Maximum out degree fraction (Max ODF) [Flake et al '00]**

$$f(S) = \max_{u \in S} \frac{|\{(u,v) \in E : v \notin S\}|}{d_u}$$

Measures the maximum fraction of edges of a node in **S** that point outside **S**



- **Average out degree fraction (Avg ODF) [Flake et al '00]**

$$f(S) = \frac{1}{n_s} \sum_{u \in S} \frac{|\{(u,v) \in E : v \notin S\}|}{d_u}$$

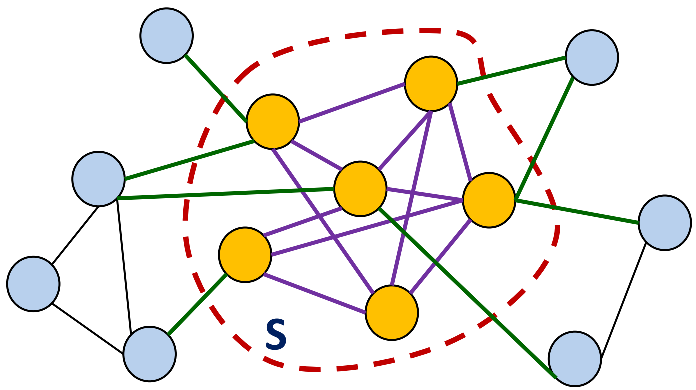
Measures the average fraction of edges of nodes in **S** that point outside **S**

Evaluation based on internal and external connectivity (3)

- **Flake's out degree fraction (Flake's ODF) [Flake et al '00]**

$$f(S) = \frac{\left|\left\{u : u \in S, \left|\left\{(u,v) \in E : v \in S\right\}\right| < d_u / 2\right\}\right|}{n_s}$$

Measures the fraction of nodes in **S** that have fewer edges pointing inside than outside of **S**



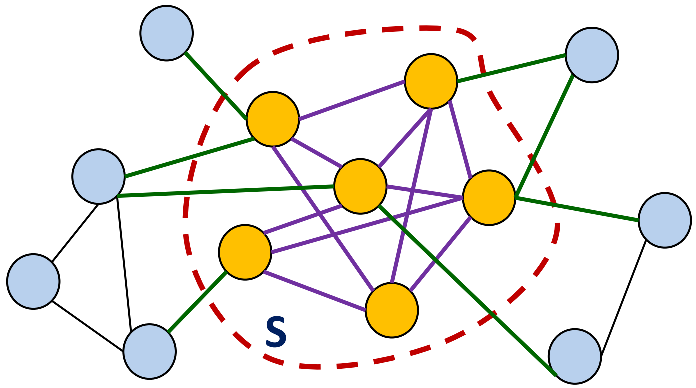
Evaluation based on network model

- **Modularity [Newman and Girvan '04], [Newman '06]**

$$f(S) = \frac{1}{4}\left(m_s - E\left(m_s\right)\right)$$

Measures the difference between the number of edges in **S** and the expected number of edges **E(m_s)** in case of a configuration model

- Typically, a random graph model with the same degree sequence



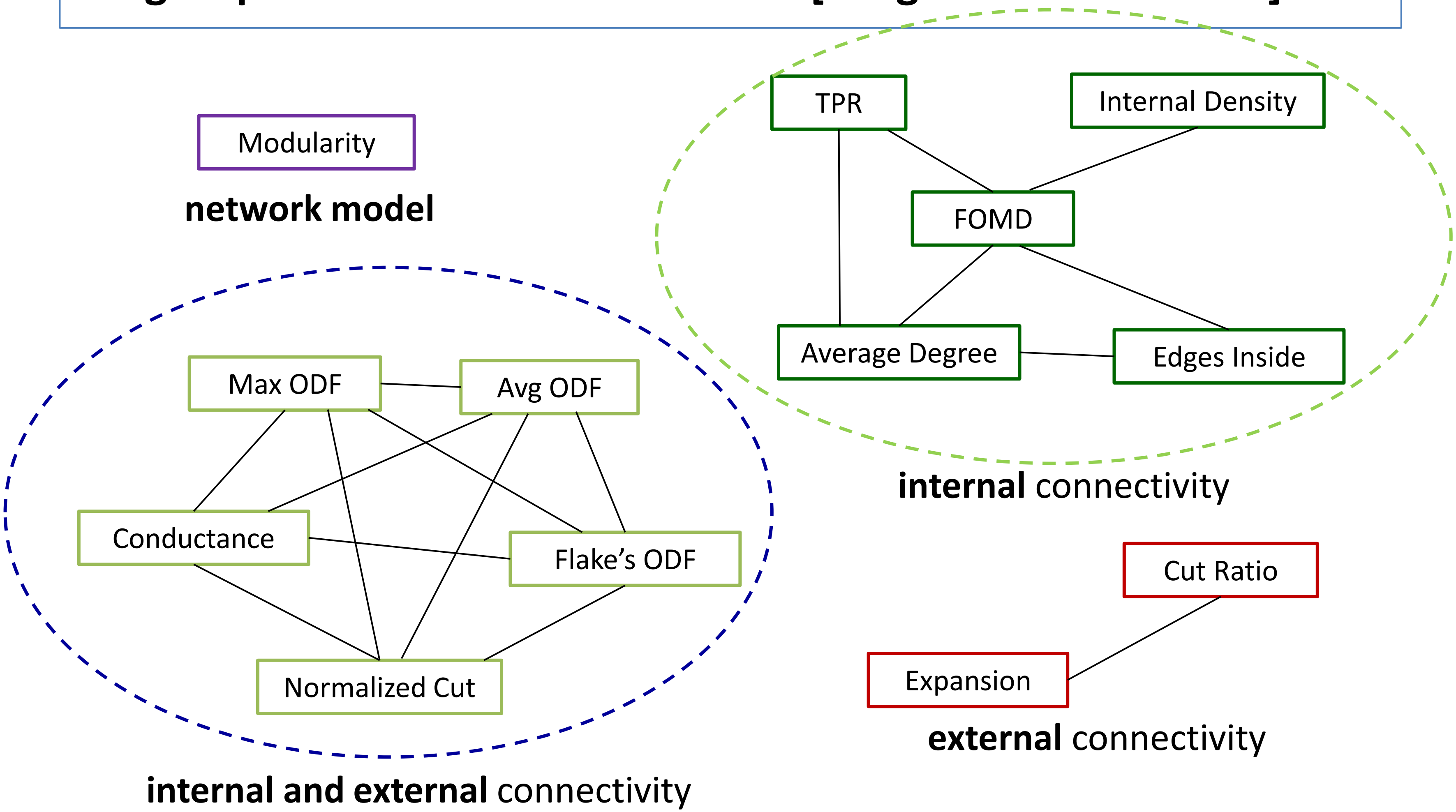
How different are the evaluation measures? (1)

- Several community evaluation measures (objective criteria) have been proposed
- Is there any **relationship** between them?
- Consider real graphs with **known node assignment to communities (ground-truth information)** and test the behavior of the objective measures **[Yang and Leskovec '12]**

1. For each of the ground-truth communities **S**
2. Compute the score of **S** using each of the previously described evaluation measures
3. Form the **correlation matrix** of the objective measures based on the scores
4. Apply a threshold in the correlation matrix
5. Extract the correlations between community objective measures

How different are the evaluation measures? (2)

- **Observation:** Community evaluation measures form **four groups** based on their correlation **[Yang and Leskovec '12]**



How different are the evaluation measures? (3)

- The different structural definitions of communities are **heavily correlated [Yang and Leskovec '12]**
- Community evaluation measures form **four groups** based on their correlation
- These groups correspond to the four main notions of structural communities
  - Communities based on **internal** connectivity
  - Communities based on **external** connectivity
  - Communities based on **internal and external** connectivity
  - Communities based on a **network model** (modularity)

References (community evaluation measures)

- M.E.J. Newman. The structure and function of complex networks. SIAM REVIEW 45, 2003.
- M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E 69(02), 2004.
- S.E. Schaeffer. Graph clustering. Computer Science Review 1(1), 2007.
- S. Fortunato. Community detection in graphs. Physics Reports 486 (3-5), 2010.
- L. Danon, J. Duch, A. Arenas, and A. Diaz-guilera. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment 9008 , 2005.
- M. Coscia, F. Giannotti, and D. Pedreschi. A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining 4 (5), 2011.
- J. Leskovec, K.J. Lang, and M.W. Mahoney. Empirical comparison of algorithms for network community detection. In: WWW, 2010.
- F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. PNAS, 101(9), 2004.
- J. Yang and J. Leskovec. Defining and Evaluating Network Communities based on Ground-Truth. In: ICDM, 2012.
- Fan Chung. Spectral Graph Theory. CBMS Lecture Notes 92, AMS Publications, 1997.

References (community evaluation measures)

- J. Shi and J. Malik. Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 2000.
- M.E.J. Newman. Modularity and community structure in networks. PNAS, 103(23), 2006.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Graph clustering algorithms

Μάθημα: Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό

Ενότητα # 5: Community Detection and Evaluation in Social and Information Networks

Διδάσκων: Μιχάλης Βαζιργιάννης

Τμήμα: Προπτυχιακό Πρόγραμμα Σπουδών “Πληροφορικής”

Graph Clustering Algorithms

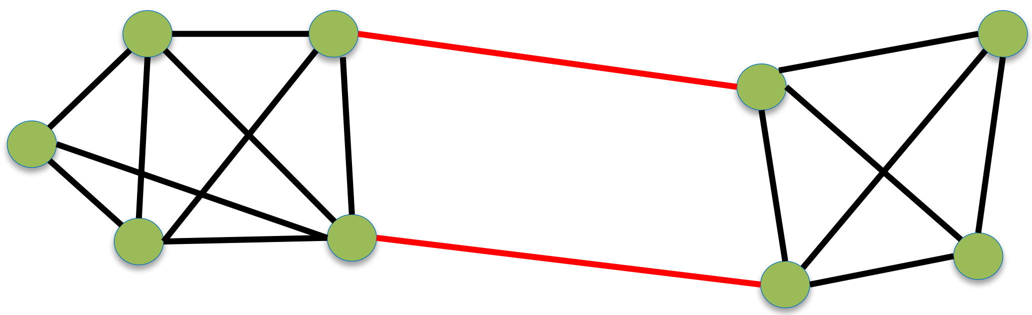
- **Spectral Clustering**
- Modularity Based Methods

Notations

- Given Graph G=(V,E) undirected:
  - Vertex Set V={v_1,.....v_n}, Edge e_{ij} between v_i and v_j
    - we assume weight w_{ij}>0 for e_{ij}
  - |V| : number of vertices
  - d_i degree of v_i : $d_i = \sum_{v_j \in V} w_{ij}$
  - $v(V) = \sum_{v_i \in V} d_i$
  - for $A \subset V$ $\bar{A} = V - A$
  - Given $\mathrm{A}, \mathrm{B} \subset V$ & $\mathrm{A} \cap B = \emptyset, w(A, B) = \sum_{v_i \in A, vj \in B} w_{ij}$
  - D : Diagonal matrix where $\mathrm{D}(\mathrm{i}, \mathrm{i}) = d_i$
  - W : Adjacency matrix $\mathrm{W}(\mathrm{i}, j) = w_{ij}$

Graph-Cut

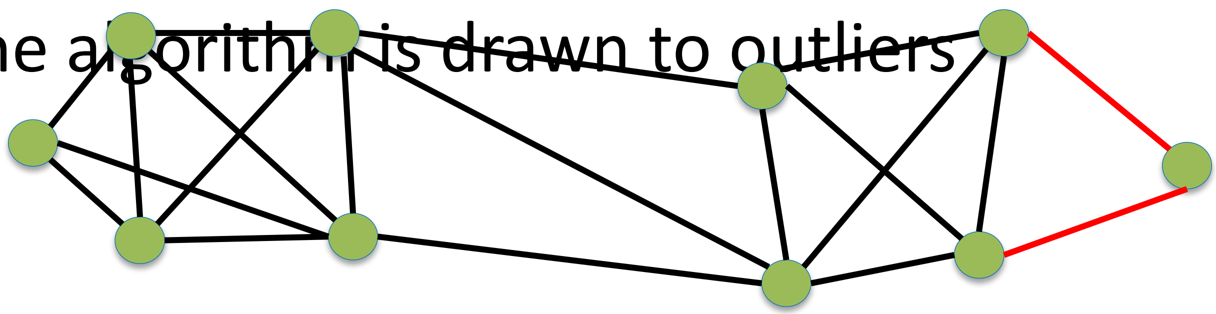
- For k clusters:
  - $cut(A_1, .., Ak) = 1/2 \sum_{i=1}^{k} w(Ai, \overline{A_i})$
    - undirected graph:1/2 we count twice each edge



- Min-cut:Minimize the edges' weight a cluster shares with the rest of the graph

Min-Cut

- Easy for k=2 : Mincut(A_1,A_2)
  - Stoer and Wagner: "A Simple Min-Cut Algorithm"
- In practice one vertex is separated from the rest
  - The algorithm is drawn to outliers



Normalized Graph Cuts

- We can normalize by the size of the cluster (size of sub-graph) :
  - number of Vertices (Hagen and Kahng, 1992):

    $Ratiocut(A_1, \dots Ak) = \sum_{i=1}^{k} \frac{cut(Ai, \overline{A_i})}{|Ai|}$
  - sum of weights (Shi and Malik, 2000) :

    $Ncut(A_1, \dots Ak) = \sum_{i=1}^{k} \frac{cut(Ai, \overline{A_i})}{v(A_i)}$
- Optimizing these functions is NP-hard
- Spectral Clustering provides solution to a relaxed version of the above

From Graph Cuts to Spectral Clustering

- For simplicity assume k=2:
  - Define $f: V \to \mathbb{R}$ for Graph G :

$$f_i = \begin{cases} 1 & v_i \in A \\ -1 & v_i \in \overline{A} \end{cases}$$

- Optimizing the original cut is equivalent to an optimization of:

$$\sum_{i,j=1}^{n} w_{ij}\left(f_i - f_j\right)^2$$

$$= \sum_{v_i \in A, v j \in \overline{A}} w_{ij}(1+1)^2 + \sum_{v_i \in \overline{A}, v j \in A} w_{ij}(-1-1)^2$$

$$= \boldsymbol{8 * cut\left(A, \overline{A}\right)}$$

Graph Laplacian

- How is the previous useful in Spectral clustering?

$$\sum_{i,j=1}^{n} w_{ij}(f_i - fj)^2$$

$$= \sum_{i,j=1}^{n} w_{ij} f_i^{\,2} \; -2 \sum_{i,j=1}^{n} w_{ij} f_i f_j + \sum_{i,j=1}^{n} w_{ij} f_j^{\,2}$$

$$= \sum_{i,j=1}^{n} d_i f_i^2 - 2 \sum_{i,j=1}^{n} w_{ij} f_i f_j + \sum_{i,j=1}^{n} d_j f_j^2$$

$$= 2\left(\sum_{i,j=1}^{n} d_{ii} f_i^2 - \sum_{i,j=1}^{n} w_{ij} f_i f_j\right)$$

$$= 2\left(\boldsymbol{f}^{\boldsymbol{T}}\boldsymbol{D}\boldsymbol{f} - \boldsymbol{f}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{f}\right) = 2\boldsymbol{f}^{\boldsymbol{T}}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{f} = 2\boldsymbol{f}^{\boldsymbol{T}}\boldsymbol{L}\boldsymbol{f}$$

- **f:**a single vector with the cluster assignments of the vertices
- **L=D-W :** the Laplacian of a graph

Properties of L

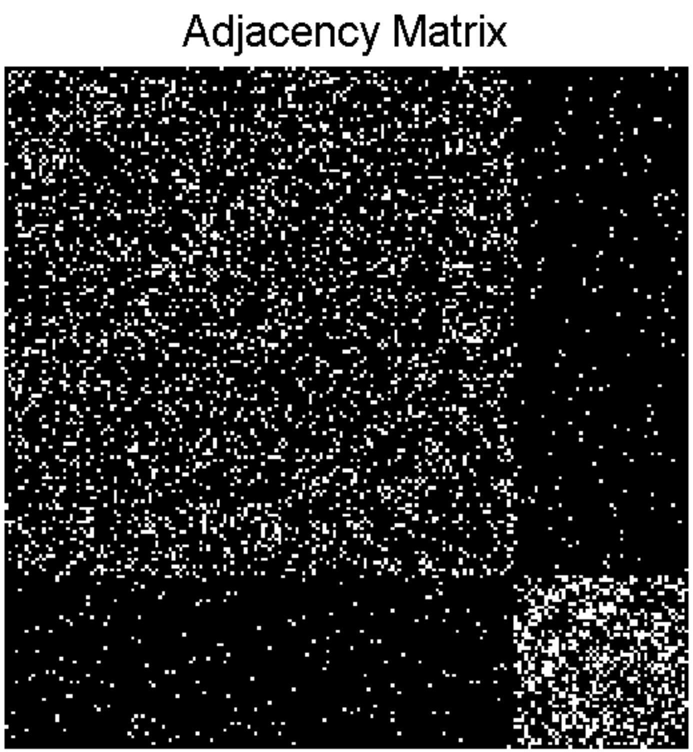
- L is
  - Symmetric
  - Positive
  - Semi-definite
- The smallest eigenvalue of L is 0
  - The corresponding eigenvector is $\mathbb{1}$
- L has n non-negative, real valued eigenvalues
  - $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$

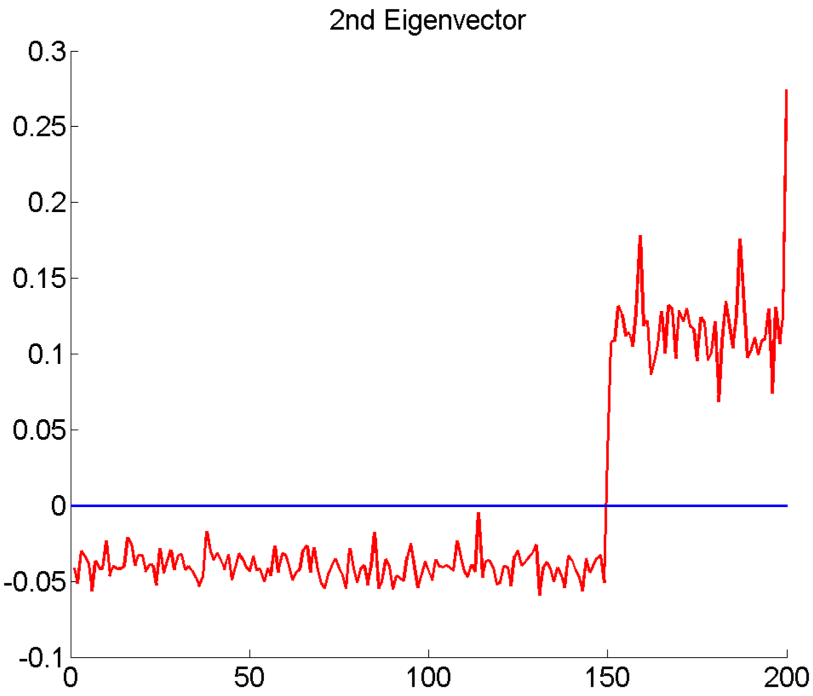
Two Way Cut from the Laplacian

- We could solve $min_f f^T Lf$ where $f \in \{-1,1\}^n$
- NP-Hard for discrete cluster assignments
  - Relax the constraint *to* $f \in R^n$:

    $min_f f^T Lf$ subject to $f^T f=n$
- The solution to this problem is given by:
  - (**Rayleigh-Ritz Theorem**) the eigenvector corresponding to smallest eigenvalue: 0 and the corresponding eigenvector (full of 1s) offers no information
- We use the second eigenvector as an approximation
  - $f_i>0$ the vertex belongs to one cluster , $fi<0$ to the other

Example


Adjacency Matrix


2nd Eigenvector
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
0
50
100
150
200

Ratio Cut

- $Ratiocut(A_1, \ldots Ak) = \sum_{i=1}^{k} \frac{cut(Ai, \overline{A}_i)}{|Ai|}$
  - Define $f: V \to \mathbb{R}$ for Graph G :

$$f_i = \begin{cases} \sqrt{\frac{|\overline{A}|}{|A|}} & vi \in A \\ -\sqrt{\frac{|A|}{|\overline{A}|}} & v_i \in \overline{A} \end{cases}$$

  - $\sum_{i,j=1}^{n} w_{ij}(f_i - fj)^2 = 2cut(\mathrm{A}, \overline{A})\left(\sqrt{\frac{|\overline{A}|}{|A|}} + \sqrt{\frac{|A|}{|\overline{A}|}} + 2\right)$
    $= 2|V|Ratiocut(A, \overline{A})$

Ratio Cut

- We have $min_f f^T L f$ subject to
  $$f^T 1 = 0,\ \ f T f = n$$

$$f^T 1 = \sum_i^n f_i = \sum_{v_i \in A} \sqrt{\frac{|\bar{A}|}{|A|}} + \sum_{v_i \in \bar{A}} -\sqrt{\frac{|A|}{|\bar{A}|}} = |\mathrm{A}| \sqrt{\frac{|\bar{A}|}{|A|}} - |\bar{A}| \sqrt{\frac{|A|}{|\bar{A}|}} = 0$$

$$f^T f = \sum_i^n f_i^2 = |\bar{A}| + |A| = n$$

- The second smallest eigenvalue of $Lf = \lambda f$ approximates the solution

Normalized Cut

- $Ncut(A_1, \dots Ak) = \sum_{i=1}^{k} \frac{cut(Ai, \overline{A}_i)}{v(A_i)}$
- Define $f : V \to \mathbb{R}$ for Graph G :

$$f_i = \begin{cases} \sqrt{\dfrac{v(\overline{A})}{v(A)}} & vi \in A \\ -\sqrt{\dfrac{v(A)}{v(\overline{A})}} & vi \in \overline{A} \end{cases}$$

- $\sum_{i,j=1}^{n} w_{ij}(f_i - fj)^2 = 2cut(\mathrm{A}, \overline{A})\left(\sqrt{\frac{v(\overline{A})}{v(A)}} + \sqrt{\frac{v(A)}{v(\overline{A})}} + 2\right)$

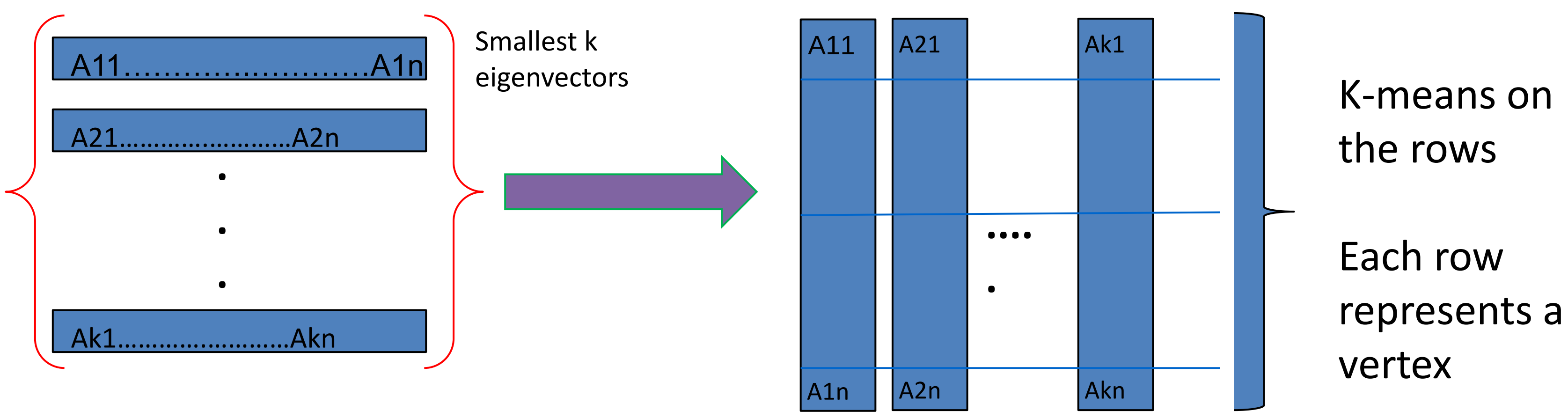
$$= 2v(\mathrm{V})\mathrm{Ncut}(\mathrm{A}, \overline{A})$$

Normalized Cut

- Similarly we come to : $min_f f^T L f$
  subject to $f^T D 1 = 0,\ fTDf = v(V)$
- Assume $h = D^{1/2} f$
  - $min_h h^T D^{-1/2} \mathrm{L} D^{-1/2} \mathrm{h}$ subject to
    $h^T D^{1/2} 1 = 0, \qquad h^T h = v(V)$
  - The answer is in the eigenvector of the second smallest eigenvalue of $L_{sym} = D^{-1/2} \mathrm{L} D^{-1/2}$
    Shi and Malik (2000)
- L_{sym} is the normalized Laplacian
  - has n non-negative, real valued eigenvalues
  - $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$

Multi-Way Graph Partition

- The cluster assignment is given by the smallest k eigenvectors of L
- The real values need to be converted to cluster assignments
  - We use k-means to cluster the rows
  - We can substitute L with L_{sym}



References

- UIrike von Luxburg, A Tutorial on Spectral Clustering, Statistics and Computing, 2007
- Davis, C., W. M. Kahan (March 1970). The rotation of eigenvectors by a perturbation. III. SIAM J. Numerical Analysis 7
- Shi, Jianbo, and Jitendra Malik. "Normalized cuts and image segmentation, "*Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2000).
- Mechthild Stoer and Frank Wagner. 1997. A simple min-cut algorithm. *J. ACM*
- Ng, Jordan & Weiss, K-means algorithm on the embedded eigen-space, NIPS 2001
- Hagen, L. Kahng, , "New spectral methods for ratio cut partitioning and clustering," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on* , 1992

Graph Clustering Algorithms

- Spectral Clustering
- **Modularity Based Methods**

Basics

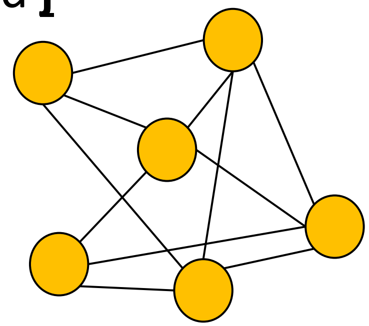
- Most of the community evaluation measures (e.g., conductance, cut-based measures), quantify the quality of a community based on
  - **Internal connectivity** (intra-community edges)
  - **External connectivity** (inter-community edges)
- **Question:** Is there any other way to distinguish groups of nodes with good community structure?
- **Random graphs** are not expected to present inherent community structure
- **Idea:** Compare the number of edges that lie **within a cluster** with the expected one in case of **random graphs** with the same degree distribution – **modularity measure**

Main idea

- **Modularity** function **[Newman and Girvan '04], [Newman '06]**
- Initially introduced as a measure for assessing the strength of communities
  - **Q = (fraction of edges within communities) –**

    (expected number of edges within communities)
- What is the **expected** number of edges?
- Consider a configuration model
  - **Random graph** model with the same degree distribution
  - Let $\mathbf{P_{ij}}$ = probability of an edge between nodes **i** and **j**

    with degrees $\mathbf{k_i}$ and $\mathbf{k_j}$ respectively
  - Then $\mathbf{P_{ij} = k_i\, k_j / 2m}$, where $\mathbf{m = |E| = \frac{1}{2} \sum_i k_i}$



Formal definition of modularity

- **Modularity Q**

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

where

- **A** is the adjacency matrix
- $\mathbf{k_i}$, $\mathbf{k_j}$ the degrees of nodes **i** and **j** respectively
- **m** is the number of edges
- $\mathbf{C_i}$ is the community of node **i**
- **δ(.)** is the Kronecker function: 1 if both nodes **i** and **j** belong on the same community ($\mathbf{C_i = C_j}$), 0 otherwise

[Newman and Girvan '04], [Newman '06]

Properties of modularity

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

- **Larger** modularity **Q** indicates **better** communities (more than random intra-cluster density)
  - The community structure would be better if the number of internal edges exceed the expected number
- Modularity value is always **smaller than 1**
- It can also take **negative values**
  - E.g., if each node is a community itself
  - No partitions with positive modularity → No community structure
  - Partitions with large negative modularity → Existence of subgraphs with small internal number of edges and large number of inter-community edges

[Newman and Girvan '04], [Newman '06], [Fortunato '10]

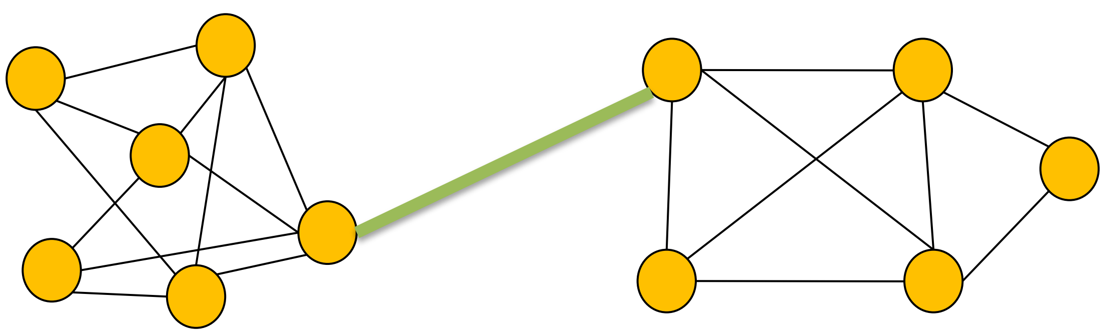
Applications of modularity

- Modularity can be applied:
  - As **quality function** in clustering algorithms
  - As **evaluation measure** for comparison of different partitions or algorithms
  - As a community detection tool itself
    - ❑ **Modularity optimization**
  - As criterion for reducing the size of a graph
    - ❑ Size reduction preserving modularity **[Arenas et al. ‘07]**

[Newman and Girvan ‘04], [Newman ‘06], [Fortunato ‘10]

Modularity-based community detection

- Modularity was first applied as a **stopping criterion** in the Newman-Girvan algorithm
- Newman-Girvan algorithm **[Newman and Girvan ‘04]**
  - A **divisive** algorithm (detect and remove edges that connect vertices of different communities)
  - **Idea:** try to identify the edges of the graph that are most between other vertices → responsible for connecting many node pairs
  - Select and remove edges based to the value of **betweenness centrality**
  - **Betweenness centrality:** number of **shortest paths** between every pair of nodes, that pass through an edge



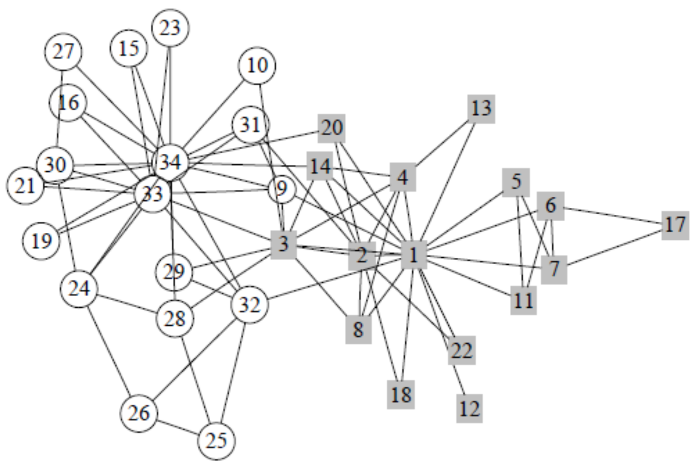
Edge betweenness is higher for edges that connect different communities

Newman-Girvan algorithm (1)

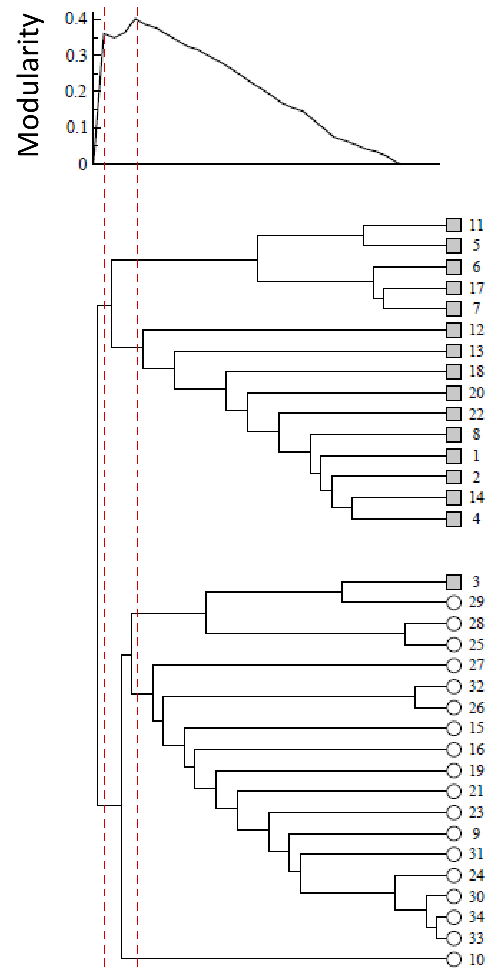
- **Basic steps:**
  1. Compute betweenness centrality for all edges in the graph
  2. Find and remove the edge with the highest score
  3. Recalculate betweenness centrality score for the remaining edges
  4. Go to step 2
- How do we know if the produced communities are **good ones** and stop the algorithm?
  - The output of the algorithm is in the form of a **dendrogram**
  - Use **modularity** as a criterion to cut the dendrogram and terminate the algorithm (Q ~= 0.3-0.7 indicates good partitions)
- Complexity: $\mathbf{O(m^2n)}$ (or $\mathbf{O(n^3)}$ on a sparse graph)

[Newman and Girvan '04], [Girvan and Newman '02]

Newman-Girvan algorithm (2)



Zachary's karate club



Community structure

[Newman and Girvan '04]

Modularity optimization

- High values of modularity indicate good quality of partitions
- **Goal:** find the partition that corresponds to the maximum value of modularity
- **Modularity maximization** problem
  - Computational difficult problem **[Brandes et al. ‘06]**
  - Appoximation techniques and heuristics
- Four main categories of techniques
  1. Greedy techniques
  2. **Spectral optimization**
  3. Simulated annealing
  4. Extremal optimization

[Fortunato ‘10]

Spectral optimization (1)

- **Idea:** Spectral techniques for modularity optimization
- **Goal:** Assign the nodes into two communities, **X** and **Y**
- Let $s_i, \forall i \in V$ be an indicator variable where $\mathbf{s_i = +1}$ if **i** is assigned to **X** and $\mathbf{s_i = -1}$ if **i** is assigned to **Y**

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

$$= \frac{1}{4m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)(s_i s_j + 1)$$

$$= \frac{1}{4m}\sum_{ij} B_{ij} s_i s_j = \frac{1}{4m} s^T \mathrm{B} s$$

- **B** is the **modularity matrix**

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

[Newman '06], [Newman '06b]

Spectral optimization (2)

- Modularity matrix **B** $$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

- Vector **s** can be written as a linear combination of the eigenvectors **u**i of the modularity matrix **B** $$s = \sum_i a_i \, u_i$$
  where $$a_i = u_i^T s$$

- Modularity can now expressed as

$$Q = \frac{1}{4m} \sum_i a_i u_i^T \, B \, \sum_j a_j u_j^T = \frac{1}{4m} \sum_{i=1}^{n} \left(u_i^T s\right)^2 \beta_i$$

Where **β**i is the eigenvalue of **B** corresponding to eigenvector **u**i

[Newman ‘06], [Newman ‘06b]

Spectral optimization (3)

- Spectral modularity optimization algorithm

  1. Consider the eigenvector $\mathbf{u_1}$ of **B** corresponding to the largest eigenvalue
  2. Assign the nodes of the graph in one of the two communities **X** ($\mathbf{s_i = +1}$) and **Y** ($\mathbf{s_i = -1}$) based on the **signs** of the corresponding components of the eigenvector

$$s_i = \begin{cases} 1 & \text{if } u_1(i) \geq 0 \\ -1 & \text{if } u_1(i) < 0 \end{cases}$$

  – More than two partitions?

  1. **Iteratively**, divide the produced partitions into two parts

  2. If at any step the split does not contribute to the modularity, leave the corresponding subgraph as is

  3. End when the entire graph has been splinted into no further divisible subgraphs

  – Complexity: $\mathbf{O(n^2 \log n)}$ for sparse graphs

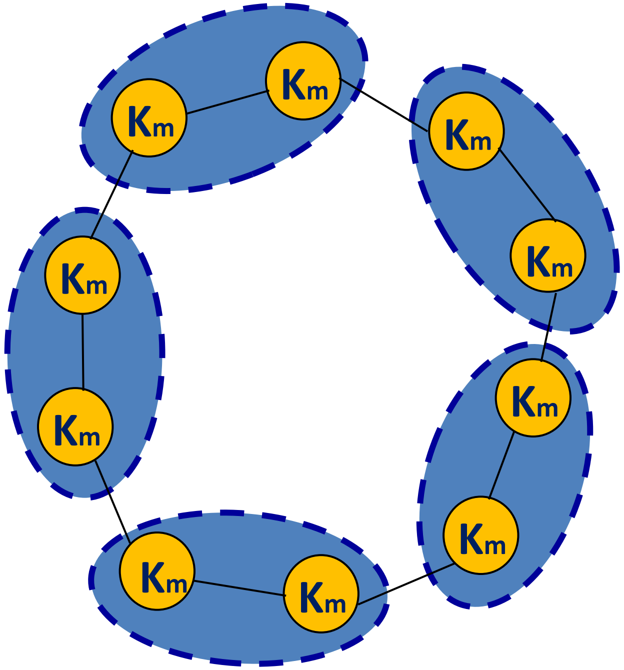
[Newman '06], [Newman '06b]

Extensions of modularity

- Modularity has been extended in several directions
  - Weighted graphs **[Newman '04]**
  - Bipartite graphs **[Guimera et al '07]**
  - Directed graphs (next in this tutorial) **[Arenas et al. '07], [Leicht and Newman '08]**
  - Overlapping community detection (next in this tutorial) **[Nicosia et al. '09]**
  - Modifications in the configuration model – local definition of modularity **[Muff et al. '05]**

Resolution limit of modularity

- **Resolution Limit** of modularity **[Fortunato and Barthelemy '07]**
- The method of modularity optimization may not detect communities with relatively small size, which depends on the total number of edges in the graph



- **K_m** are cliques with **m** edges (**m ≤ sqrt(|E|)**)
- **K_m** represent well-defined clusters
- However, the maximum modularity corresponds to clusters formed by two or more cliques
- It is difficult to know if the community returned by modularity optimization corresponds to a **single community** or a **union of smaller communities**

References (modularity)

- M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E 69(02), 2004.
- M.E.J. Newman. Modularity and community structure in networks. PNAS, 103(23), 2006.
- S.E. Schaeffer. Graph clustering. Computer Science Review 1(1), 2007.
- S. Fortunato. Community detection in graphs. Physics Reports 486 (3-5), 2010.
- M. Coscia, F. Giannotti, and D. Pedreschi. A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining 4 (5), 2011.
- A. Arenas, J. Duch, A. Fernandez, and S. Gomez. Size reduction of complex networks preserving modularity. New J. Phys., 9(176), 2007.
- M. Girvan and M.E.J. Newman. Community structure in social and biological networks. PNAS 99(12), 2002.
- U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On Modularity Clustering. IEEE TKDE 20(2), 2008.
- M.E.J. Newman. Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 2004.
- A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. Phys. Rev. E 70, 2004.

References (modularity)

- M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 2006.
- R. Guimera, M. Sales-Pardo, L.A.N. Amaral. Modularity from Fluctuations in Random Graphs and Complex Networks. Phys. Rev. E 70, 2004.
- J. Duch and A. Arenas. Community detection in complex networks using Extremal Optimization. Phys. Rev. E 72, 2005.
- A. Arenas, J. Duch, A. Fernandez, and S. Gomez. Size reduction of complex networks preserving modularity. New Journal of Physics 9(6), 2007.
- E.A. Leicht and M.E.J. Newman. Community structure in directed networks. Phys. Rev. Lett. 100, 2008.
- V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. Extending the definition of modularity to directed graphs with overlapping communities. J. Stat. Mech. 03, 2009.
- S. Muff, F. Rao, A. Caflisch. Local modularity measure for network clusterizations. Phys. Rev. E, 72, 2005.
- S. Fortunato and M. Barthelemy. Resolution limit in community detection. PNAS 104(1), 2007.

Alternative Methods for Community Evaluation

Μάθημα: Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό

Ενότητα # 5: Community Detection and Evaluation in Social and Information Networks

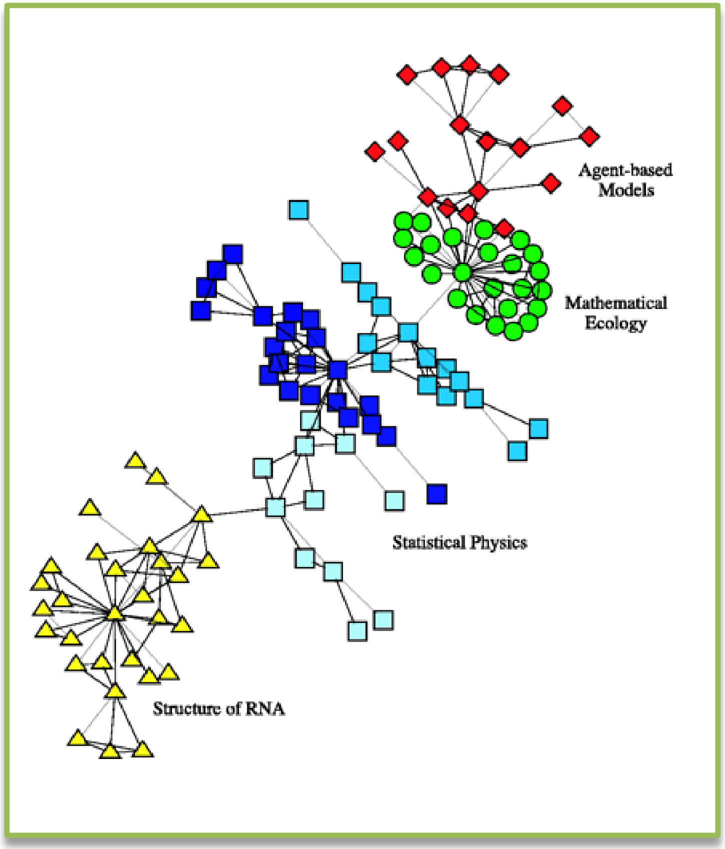
Διδάσκων: Μιχάλης Βαζιργιάννης

Τμήμα: Προπτυχιακό Πρόγραμμα Σπουδών “Πληροφορικής”

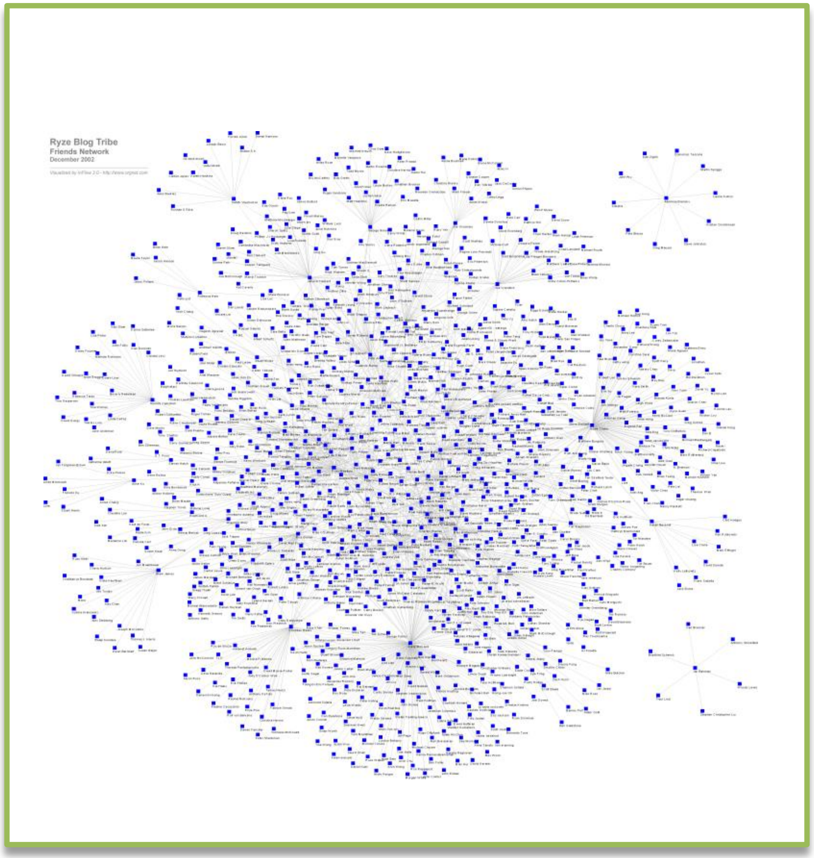
Topics on community detection and evaluation

- **Observations on structural properties of large graphs**
- Degeneracy-based community evaluation

Community structure in small vs. large graphs



Small scale collaboration network (Newman)

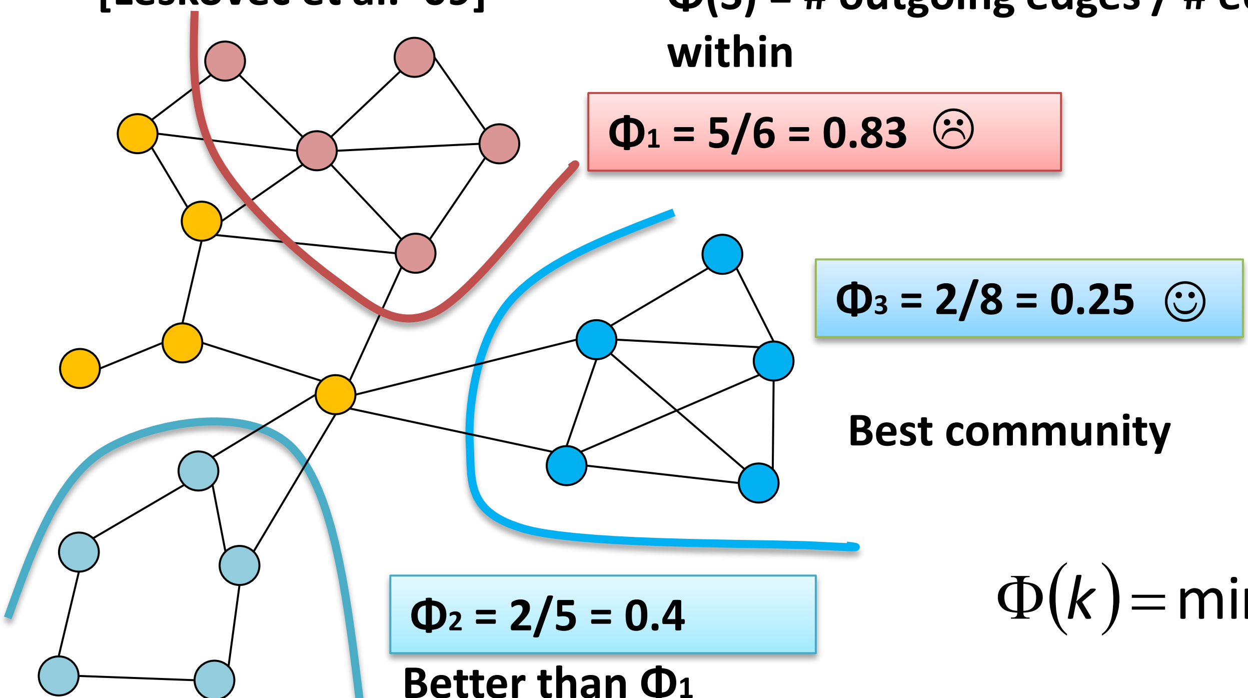


Blog network
http://www.ryze.com

Examine the structural differences

- How can we examine and compare the structural differences – **in terms of community structure** – at different scale graphs?
- Use **conductance Φ(S)** as a community evaluation measure
  - Smaller value for conductance implies better community-like properties **[Leskovec et al. ’09]**

Φ(S) = # outgoing edges / # edges within



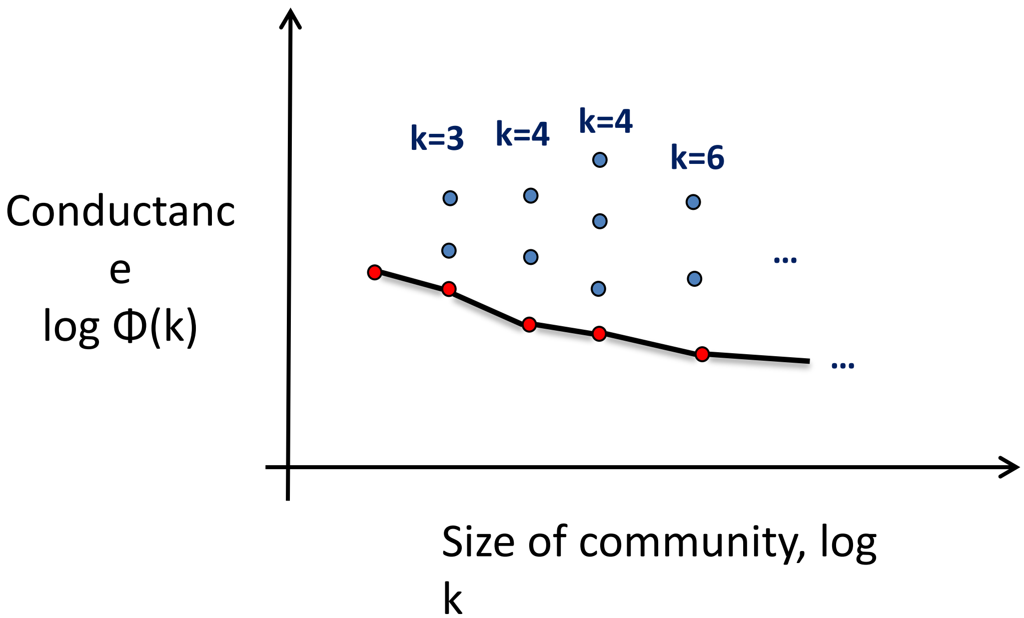
Find the best community of 5 nodes

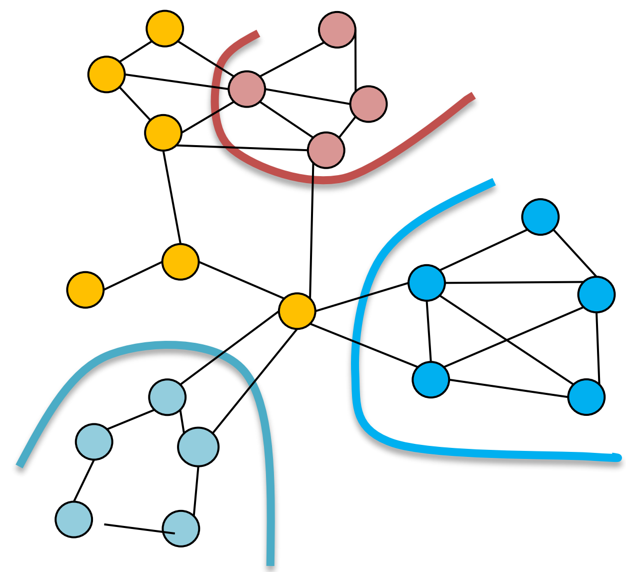
$$\Phi(k) = \min_{S \subset V, |S| = k} \Phi(S)$$

Example by J. Leskovec, ICML 2009

Network Community Profile plot

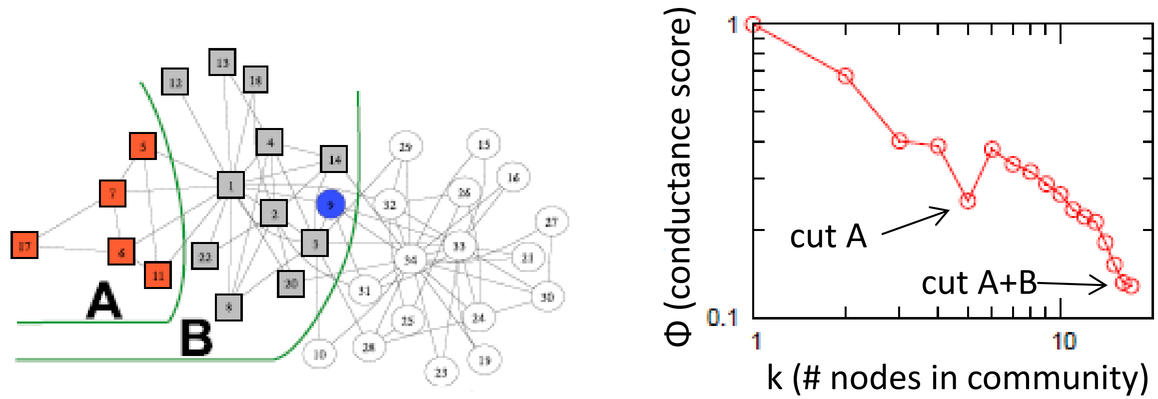
- **Network Community Profile** (NCP) plot **[Leskovec et al. '09]**
  – Plot the best conductance score (minimum) **Φ(k)** for each community size **k**





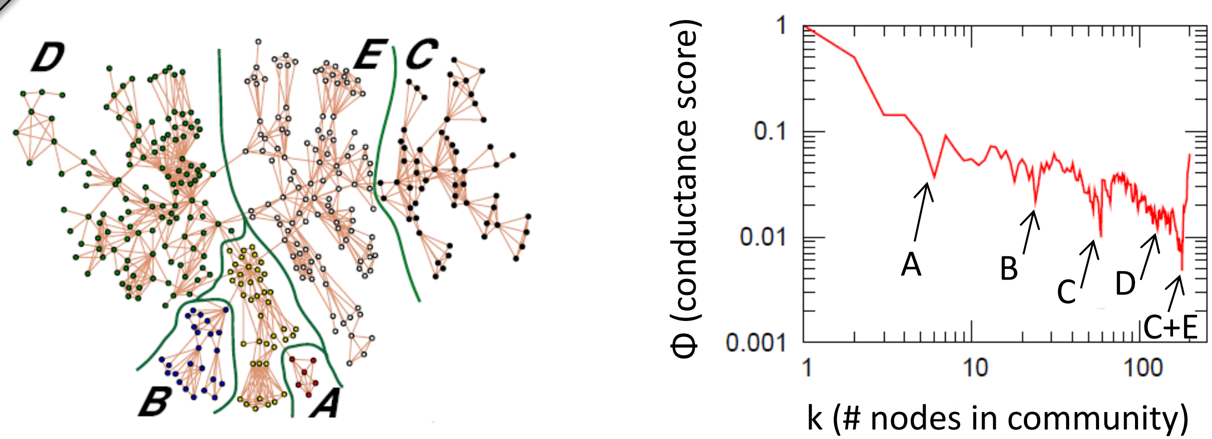
NCP plot of real graphs ?

NCP plot examples



Zachary's karate club social network

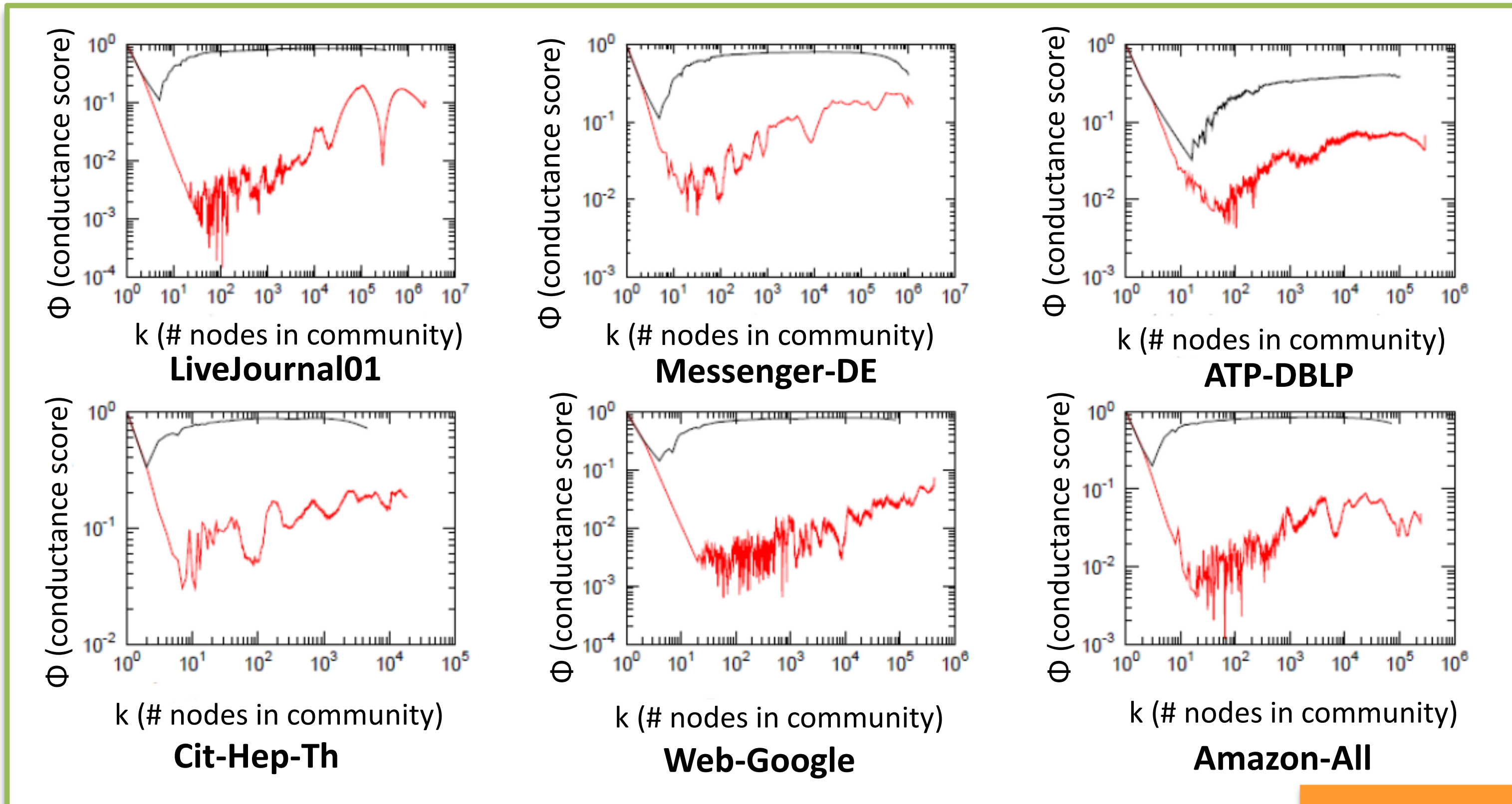
Small scale networks



Newman's collaboration network

[Leskovec et al. '09]

NCP plot of large real-world graphs



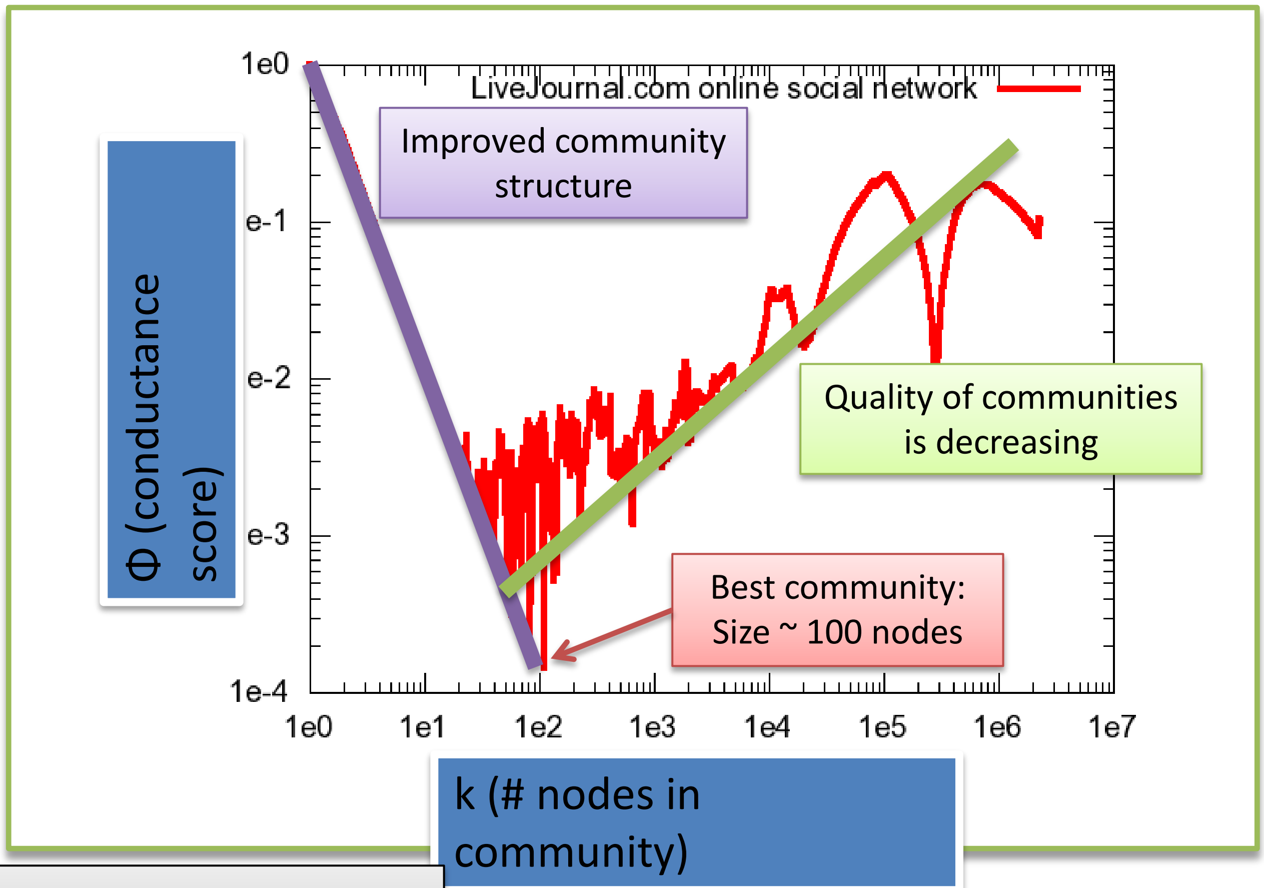
Large scale networks

Any common property?

[Leskovec et al. '09]

Figure: J. Leskovec, ICML 2009

NCP plot: Observation in large graphs



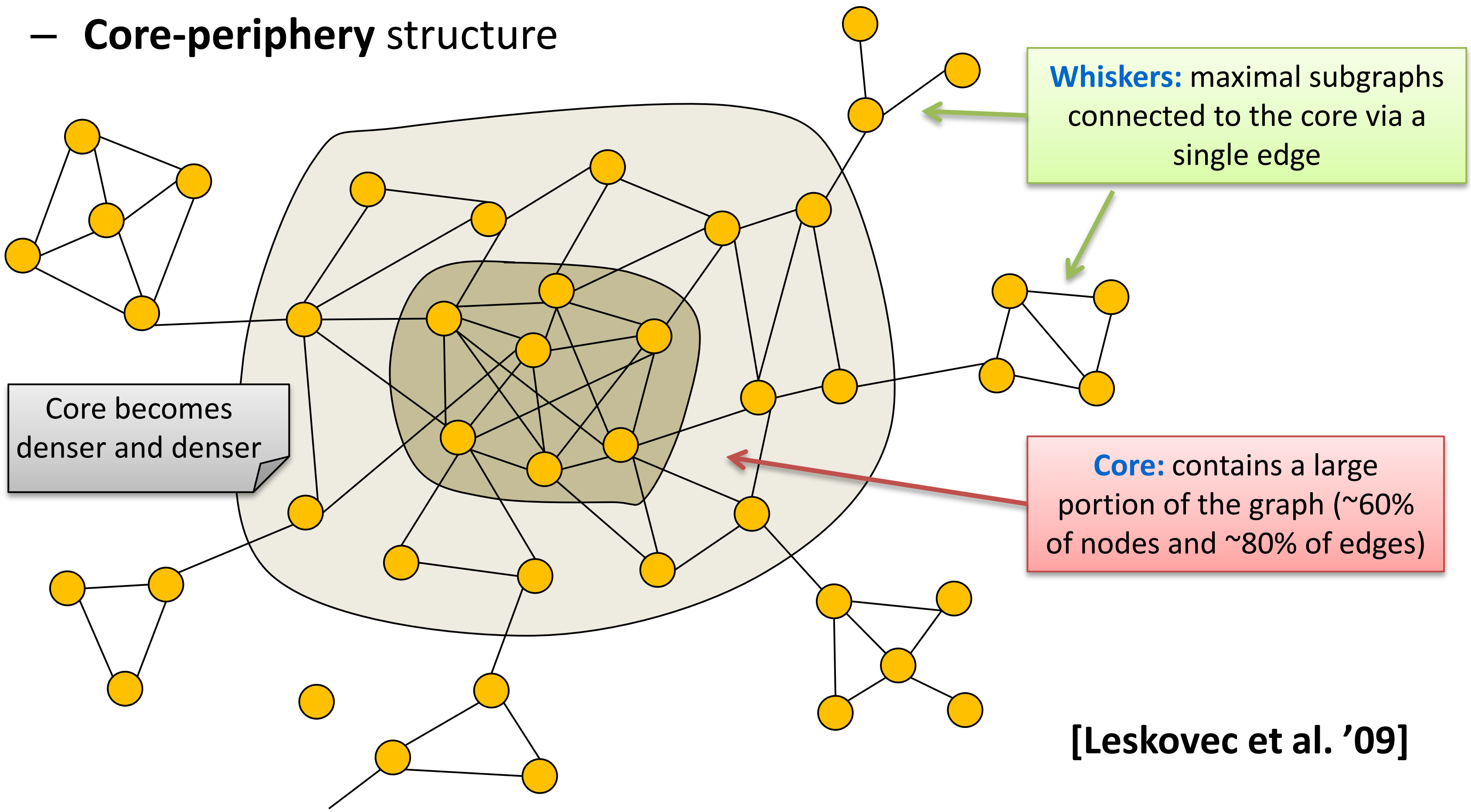
LiveJournal social network
|V| = 5M, |E| = 42 M

Figure: http://snap.stanford.edu/ncp/
Slide by J. Leskovec, ICML 2009

[Leskovec et al. '09]

Explanation: Core-Periphery structure

- How can we explain the observed structure of large graphs?
  - **Core-periphery** structure



[Leskovec et al. '09]

Similar structural observations

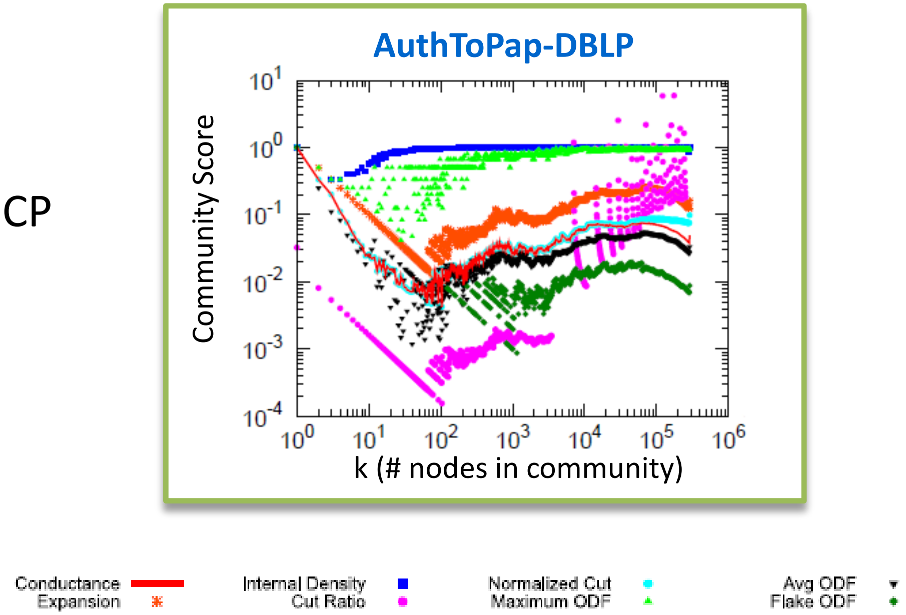
- **Jellyfish model** for the Internet topology **[Tauro et al. ’01]**

- **Min-cut** plots **[Chakrabarty et al. ’04]**
  - Perform min-cut recursively
  - Plot the relative size of the minimum cut

- **Robustness** of large scale social networks **[Malliaros et al. ’12]**
  - Robustness estimation based on the expansion properties of graphs
  - Social networks are expected to show **low robustness** due to the existence of communities → the (small number of) inter-community edges will act as bottlenecks
  - Large scale social graphs tend to be extremely robust
  - **Structural differences** (in terms of robustness and community structure) between **different scale graphs**

Clustering algorithms and objective criteria

- **Question 1:** Is the observed property an effect of the used community detection algorithm (Metis + flow based method)?
  - **A:** No. The qualitative shape of the NCP plot is the same, regardless of the community detection algorithm **[Leskovec et al. '09]**
- **Question 2:** Is the observed property an effect of the **conductance** community evaluation measure?
  - **A:** No. All the objective criteria that based on both internal and external connectivity, show a qualitatively almost similar behavior **[Leskovec et al. '10]**
  - A ***V***-like slope in the NCP plot



Conclusions

- **Large scale real-world graphs**
  - Core-periphery structure
  - No large, well defined communities
  - Structural differences between different scale graphs
- Community detection algorithms should take into account these structural observation
  - Whiskers correspond to the best (conductance-based) communities
  - Need larger high-quality clusters?
  - **Bag of whiskers:** union of disjoint (disconnected) whiskers are mainly responsible for the best high-quality clusters of larger size (above 100)

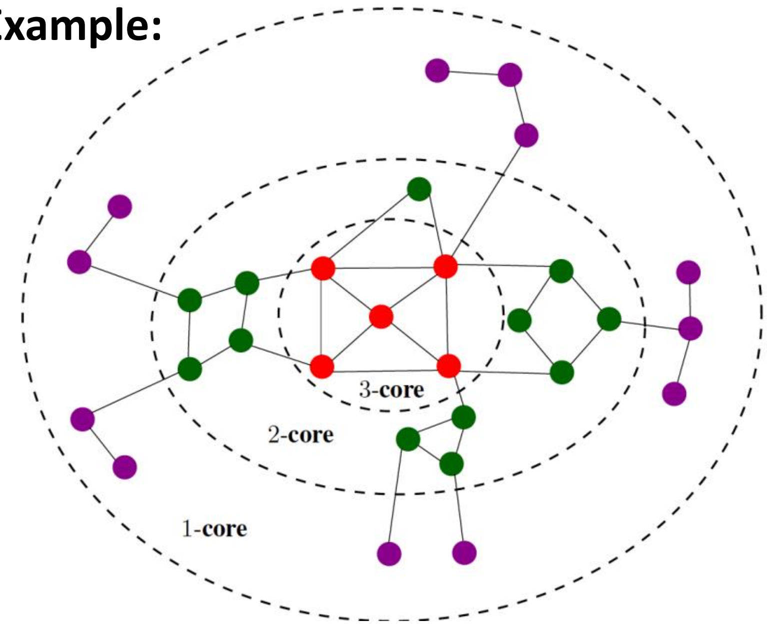
Topics on community detection and evaluation

- Observations on structural properties of large graphs
- **Degeneracy-based community evaluation**

Graph Degeneracy and the k-core Decomposition

- Degeneracy for an **undirected** graph G
  - Also known as the **k-core** number
  - The k-core of G is the largest subgraph in which every vertex has degree at least k within the subgraph

Example:



Core number $c_i = 1$

Core number $c_i = 2$

Core number $c_i = 3$

Graph Degeneracy $\delta^*(G) = 3$

$G_0 = G$

$G_1 = 1\text{-core of } G$

$G_2 = 2\text{-core of } G$

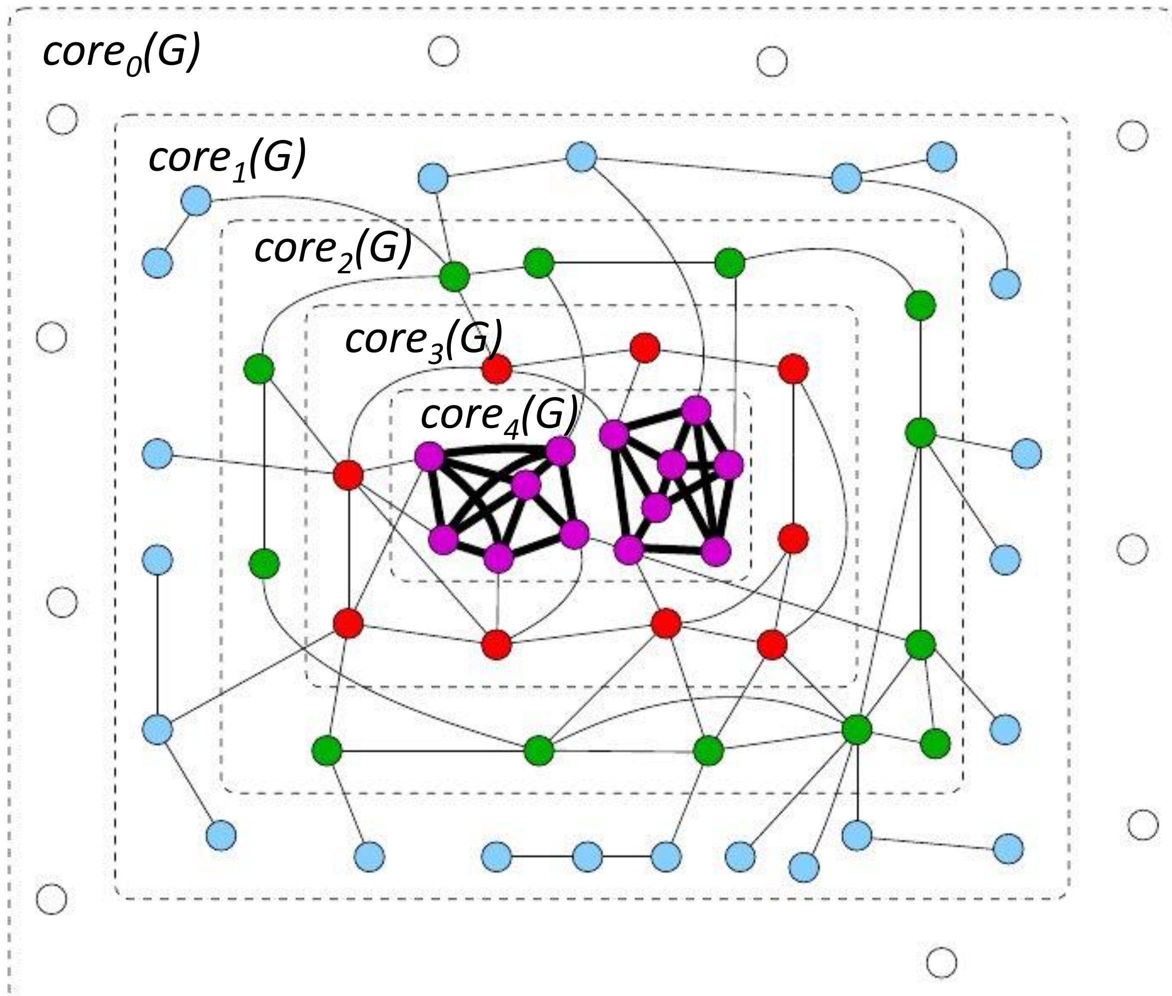
$G_3 = 3\text{-core of } G$

$G_0 \supseteq G_1 \supseteq G_2 \supseteq G_3$

Important property:
- **Fast** and easy to compute
- Almost linear to the size of the graph
- Scalable to large scale graphs

Note:
The degeneracy and the size of the k-core provide a good indication of the cohesiveness of the graph

Another example



K-core

- An algorithm for computing the *k-th core of a graph:*

  Procedure *Trim*$_k$(G, *k*)

  Input: An undirected graph G *and positive integer k*

  *Output: k**-core**(*G*)*

  1. let F := G.

  2. ***while*** *there is a node x in F such that* ***deg***$_F$*(x)<k*

      delete *node x from F.*

  3. ***return*** *F.*

- Many efficient algorithms have been given for the computation*:*
  - E.g. [*Batagelj and Zaversnik, 2003*]
    - *Time complexity: O(*m*) (*m= |E|*)*
- Fast! especially in real word data where G is usually sparse.

DBLP K-cores

- Extreme k-core: k=15 (DBLP), 76 authors
- Author ranking metric: max(k)-core that an author belongs to
  - e.g. Paul Erdos : 14
- On the max(k)-core we can identify the “closest” collaborators**: Hop-1 community**
  - **Erdos hop-1**:
    Boris Aronov, Daniel J. Kleitman, János Pach, Leonard J. Schulman, Nathan Linial, Béla Bollobás, Miklós Ajtai, Endre Szemerédi, Joel Spencer,Fan R. K. Chung, Ronald L. Graham, David Avis, Noga Alon, László Lovász, Shlomo Moran, Richard Pollack, Michael E. Saks, Shmuel Zaks, Peter Winkler, Prasad Tetali, László Babai

Degeneracy on directed graphs

(k,l)-D-*core* (G): the (k,l) D-core of graph G

for each k,l : $dc_{k,l}$ = |(k,l)-D-*core* (G)|

D-core matrix: $D(k,l)=dc_{k,l}$, k,l integers – each cell stores the size of the respective D-core

Frontier: F(D) = {(k;l): $dc_{k,l} > 0$ & $dc_{k+1,l+1} = 0$ } : the extreme (k,l)-D-cores

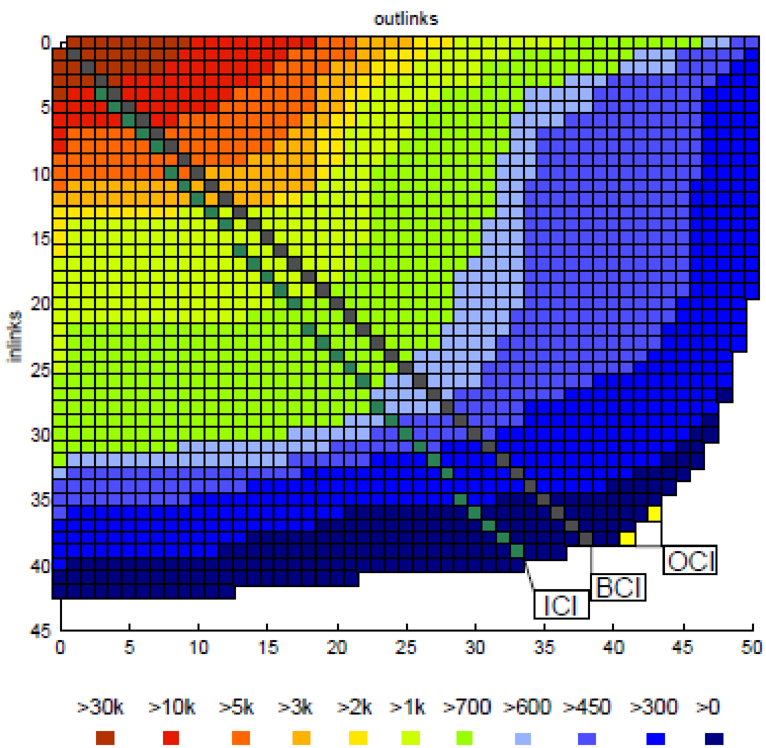
Collaboration indices

- Balanced collaboration index (**BCI**) : Intersection of diagonal D(k,k) with frontier
- Optimal collaboration index (**OCI**) : DC(k,l) where max(($k+l$)/2) distance from D($0,0$)
- Inherent collaboration index (**ICI**): All cores on the angle defined by the average inlinks/outlinks ratio

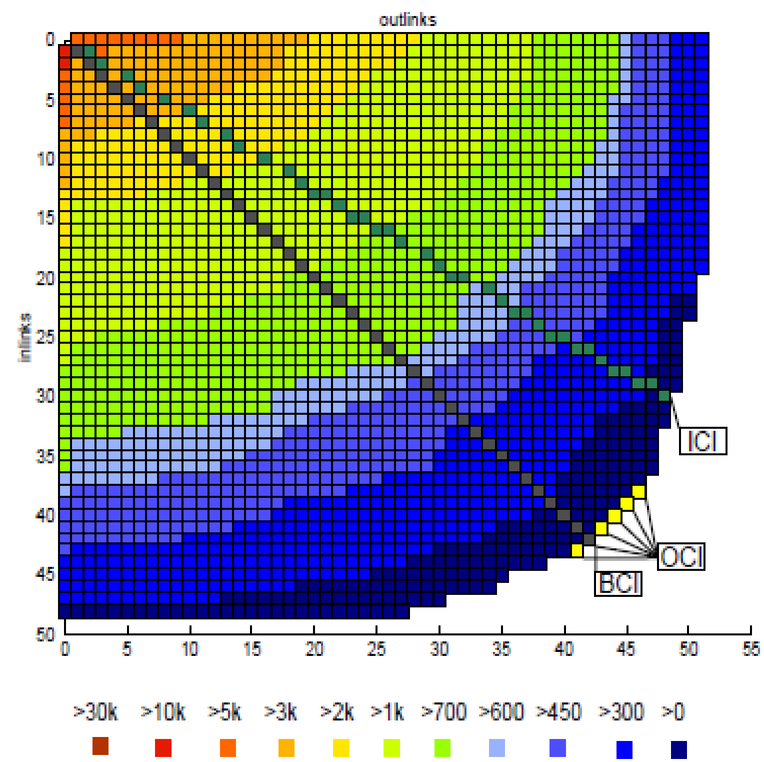
D-core matrix Wikipedia & DBLP

Extend the notion of degeneracy in directed graphs: (k, l)-D-Core

WIKIPEDIA 2004



DBLP – CITATION graph



Christos Giatsidis, Dimitrios M. Thilikos, Michalis Vazirgiannis: “D-cores: Measuring Collaboration of Directed Graphs Based on Degeneracy”, IEEE - ICDM 2011: 201-210

The Extreme DBLP D-core Authors

Authoritative and Collaborative Scientists

José A. Blakeley
Hector Garcia-Molina
Abraham Silberschatz
Umeshwar Dayal
Eric N. Hanson
Jennifer Widom
Klaus R. Dittrich
Nathan Goodman
Won Kim
Alfons Kemper
Guido Moerkotte
Clement T. Yu
M. Tamer Ã Zsu
Amit P. Sheth
Ming-Chien Shan
Richard T. Snodgrass
David Maier
Michael J. Carey
David J. DeWitt
Joel E. Richardson
Eugene J. Shekita
Waqar Hasan
Marie-Anne Neimat
Darrell Woelk
Roger King
Stanley B. Zdonik
Lawrence A. Rowe
Michael Stonebraker
Serge Abiteboul
Richard Hull
Victor Vianu
Jeffrey D. Ullman
Michael Kifer
Philip A. Bernstein
Vassos Hadzilacos
Elisa Bertino
Stefano Ceri
Georges Gardarin

Patrick Valduriez
Ramez Elmasri
Richard R. Muntz
David B. Lomet
Betty Salzberg
Shamkant B. Navathe
Arie Segev
Gio Wiederhold
Witold Litwin
Theo Härder
François Bancilhon
Raghu Ramakrishnan
Michael J. Franklin
Yannis E. Ioannidis
Henry F. Korth
S. Sudarshan
Patrick E. O'Neil
Dennis Shasha
Shamim A. Naqvi
Shalom Tsur
Christos H. Papadimitriou
Georg Lausen
Gerhard Weikum
Kotagiri Ramamohanarao
Maurizio Lenzerini
Domenico Saccà
Giuseppe Pelagatti
Paris C. Kanellakis
Jeffrey Scott Vitter
Letizia Tanca
Sophie Cluet
Timos K. Sellis
Alberto O. Mendelzon
Dennis McLeod
Calton Pu
C. Mohan
Malcolm P. Atkinson
Doron Rotem

Michel E. Adiba
Kyuseok Shim
Goetz Graefe
Jiawei Han
Edward Sciore
Rakesh Agrawal
Carlo Zaniolo
V. S. Subrahmanian
Claude Delobel
Christophe Lecluse
Michel Scholl
Peter C. Lockemann
Peter M. Schwarz
Laura M. Haas
Arnon Rosenthal
Erich J. Neuhold
Hans-Jorg Schek
Dirk Van Gucht
Hamid Pirahesh
Marc H. Scholl
Peter M. G. Apers
Allen Van Gelder
Tomasz Imielinski
Yehoshua Sagiv
Narain H. Gehani
H. V. Jagadish
Eric Simon
Peter Buneman
Dan Suciu
Christos Faloutsos
Donald D. Chamberlin
Setrag Khoshafian
Toby J. Teorey
Randy H. Katz
Miron Livny
Philip S. Yu
Stanley Y. W. Su
Henk M. Blanken

Peter Pistor
Matthias Jarke
Moshe Y. Vardi
Daniel BarbarÃ¡
Uwe Deppisch
H.-Bernhard Paul
Don S. Batory
Marco A. Casanova
Joachim W. Schmidt
Guy M. Lohman
Bruce G. Lindsay
Paul F. Wilms
Z. Meral Özsoyoglu
Gultekin Özsoyoglu
Kyu-Young Whang
Shahram Ghandeharizadeh
Tova Milo
Alon Y. Levy
Georg Gottlob
Johann Christoph Freytag
Klaus Küspert
Louiqa Raschid
John Mylopoulos
Alexander Borgida
Anand Rajaraman
Joseph M. Hellerstein
Masaru Kitsuregawa
Sumit Ganguly
Rudolf Bayer
Raymond T. Ng
Daniela Florescu
Per-Åke Larson
Hongjun Lu
Ravi Krishnamurthy
Arthur M. Keller
Catriel Beeri
Inderpal Singh Mumick
Oded Shmueli

George P. Copeland
Peter Dadam
Susan B. Davidson
Donald Kossmann
Christophe de Maindreville
Yannis Papakonstantinou
Kenneth C. Sevcik
Gabriel M. Kuper
Peter J. Haas
Jeffrey F. Naughton
Nick Roussopoulos
Bernhard Seeger
Georg Walch
R. Erbe
Balakrishna R. Iyer
Ashish Gupta
Praveen Seshadri
Walter Chang
Surajit Chaudhuri
Divesh Srivastava
Kenneth A. Ross
Arun N. Swami
Donovan A. Schneider
S. Seshadri
Edward L. Wimmers
Kenneth Salem
Scott L. Vandenberg
Dallan Quass
Michael V. Mannino
John McPherson
Shaul Dar
Sheldon J. Finkelstein
Leonard D. Shapiro
Anant Jhingran
George Lapis

Degeneracy in Signed Graphs

- Signed graphs can depict a wide variety of concepts
  - Positive/negative interactions among individuals
  - Common behavior in product review websites (e.g., epinions.com)

- A member of a directed signed graph G can either trust or distrust another but not both simultaneously

- Each vertex **v** has both positive & negative in-degree and both positive & negative out-degree

- **Our solution:** we define and extend the degeneracy concept upon a trust network

S-cores Stucture

- We compute the trust network degeneracy along the 4 combinations of **direction** and **sign** (in, out):
  - **(+,+)**: Mutual Trust
  - **(+,-)**: Trust under distrust (i.e., trust those who do not trust me)
  - **(-,-)**: Mutual distrust
  - **(-,+)**: Distrust under trust

Data Statistics

Explicit

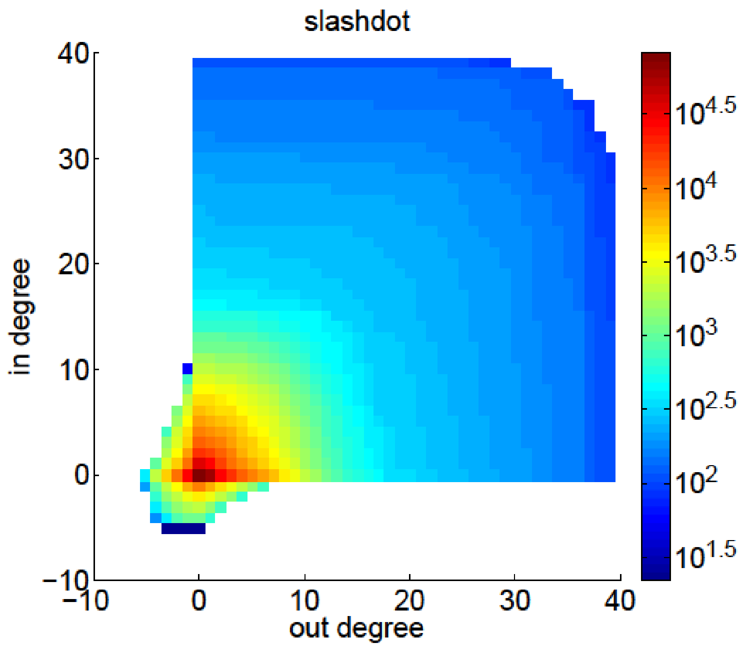
Network	Nodes	Edges	Negative
Epinions	119,217	841,200	15.0%
Slashdot	82,144	549,202	22.6%

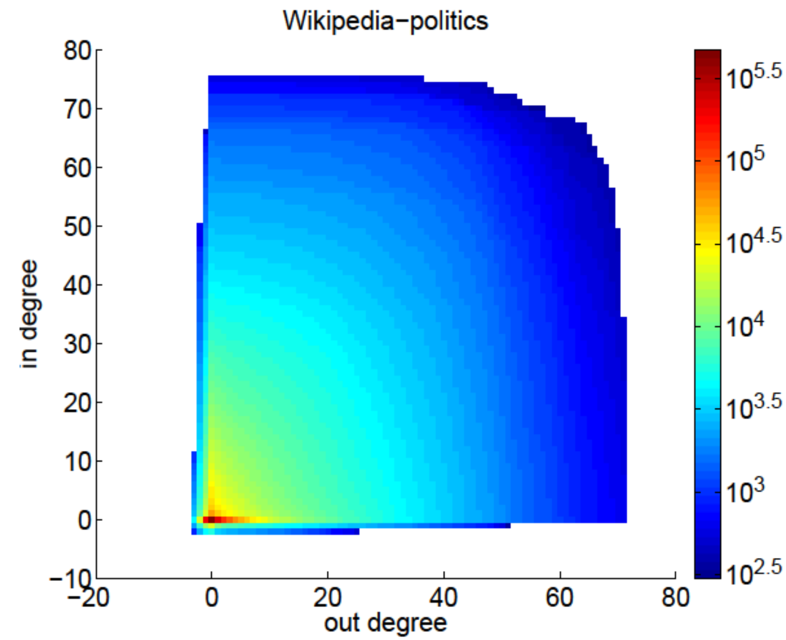
Implicit (Wikipedia)

Domain	Articles	Nodes	Edges	Positive	Negative
History	3,331	141,983	534,693	439,193	95,500
Politics	12,921	453,116	2,428,945	2,099,410	329,535
Religion	6,459	277,482	1,423,279	1,244,166	179,113
Mathematics	9,610	158,671	651,450	548,073	103,377

Examples

- **S-Cores sizes on real world data**

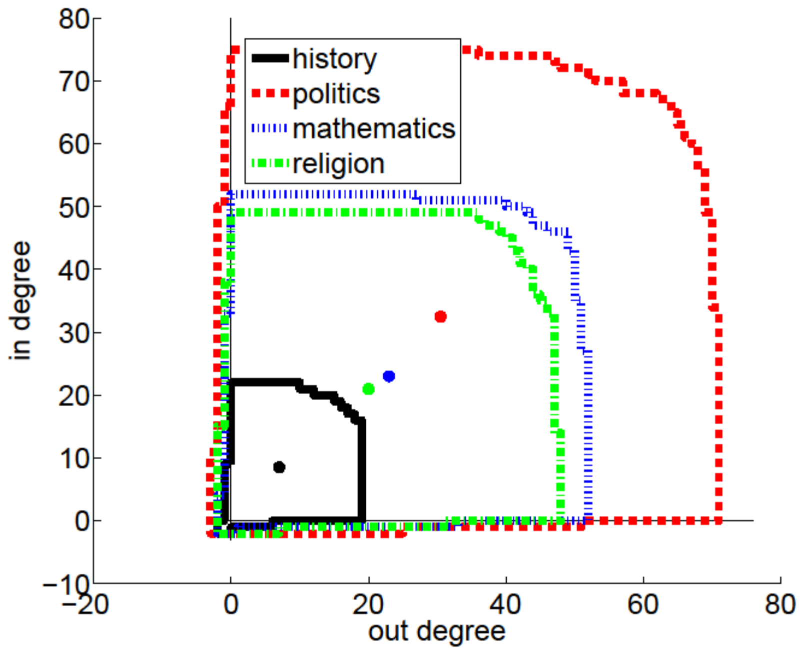




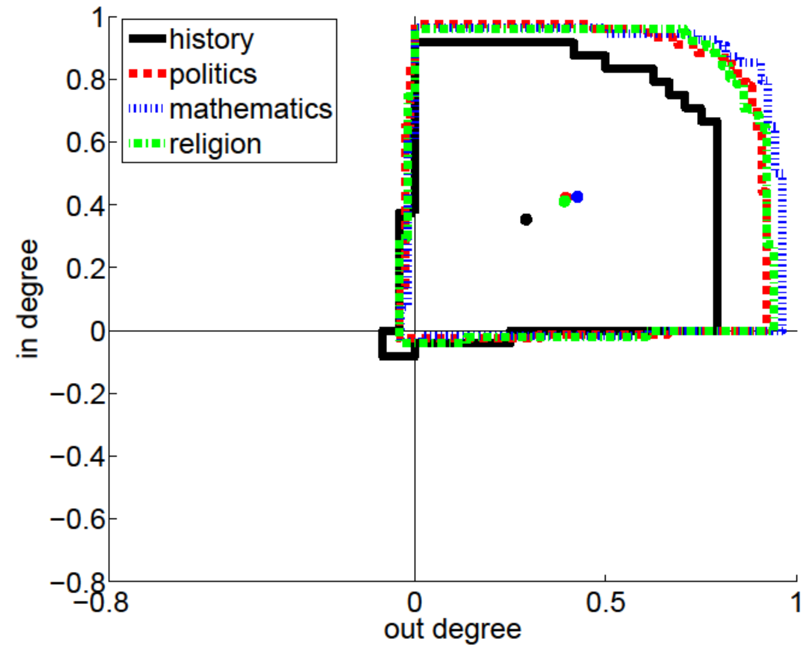
Observations:

- In both cases positive trust dominates
- In slashdot there is proportionaly much more mutual distrust than in the wikipedia-politics case

Evaluate Wikipedia Topics



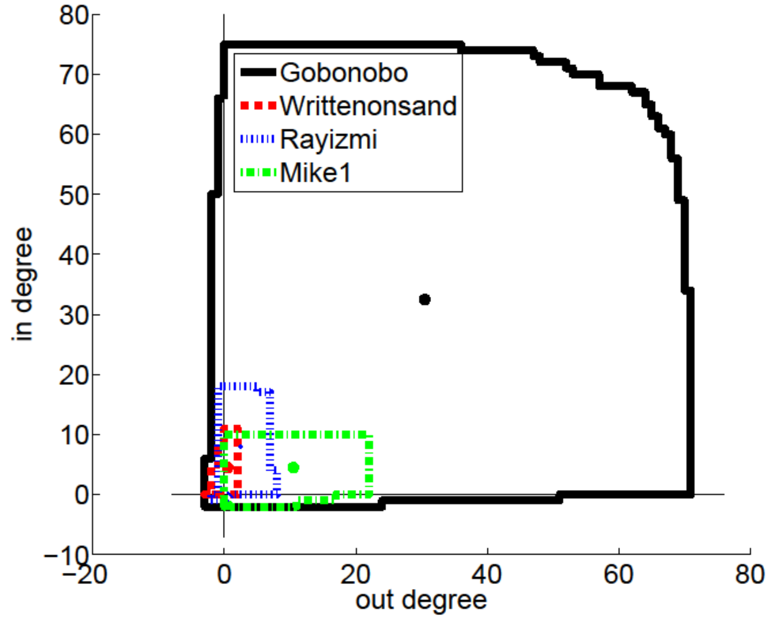
Original frontiers

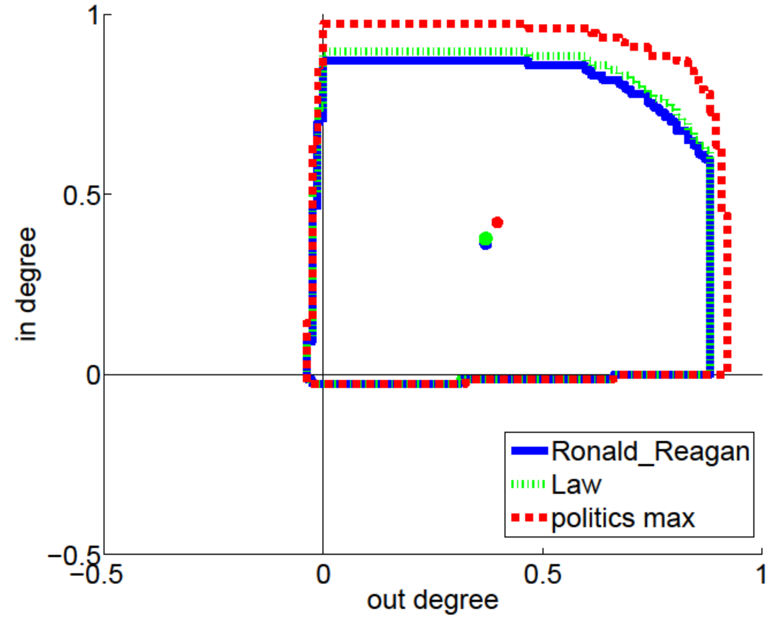


Normalized

- Wikipedia *politics* is the most robust trust network, *history* is the least one
- In the normalized case: *history* is the one with the largest mutually negative trust constituent

Users & Articles



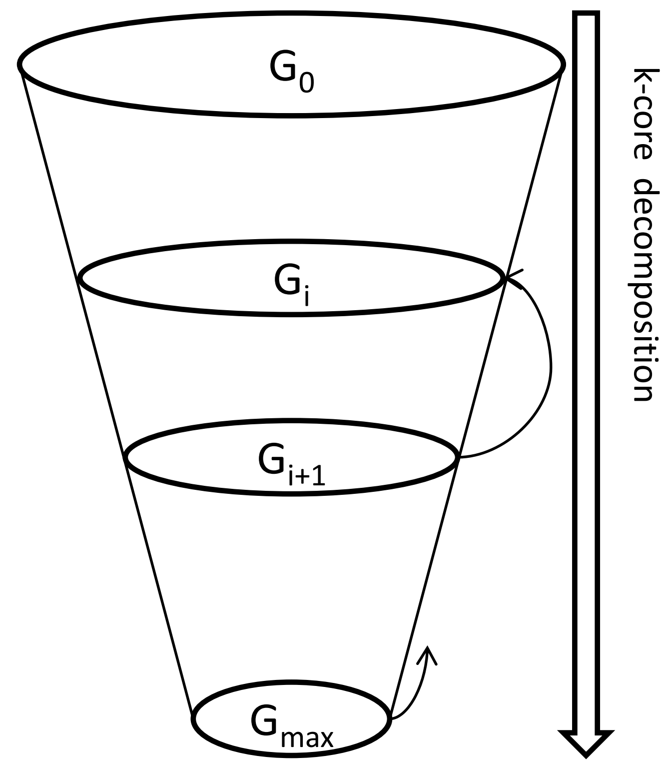


- **Editors**
  - *Gobonobo* is by far the most trusting and trusted one – i.e., a very senior editor
- **Article frontier**
  - “Reagan” article is almost as trusted as the “Politics” topic

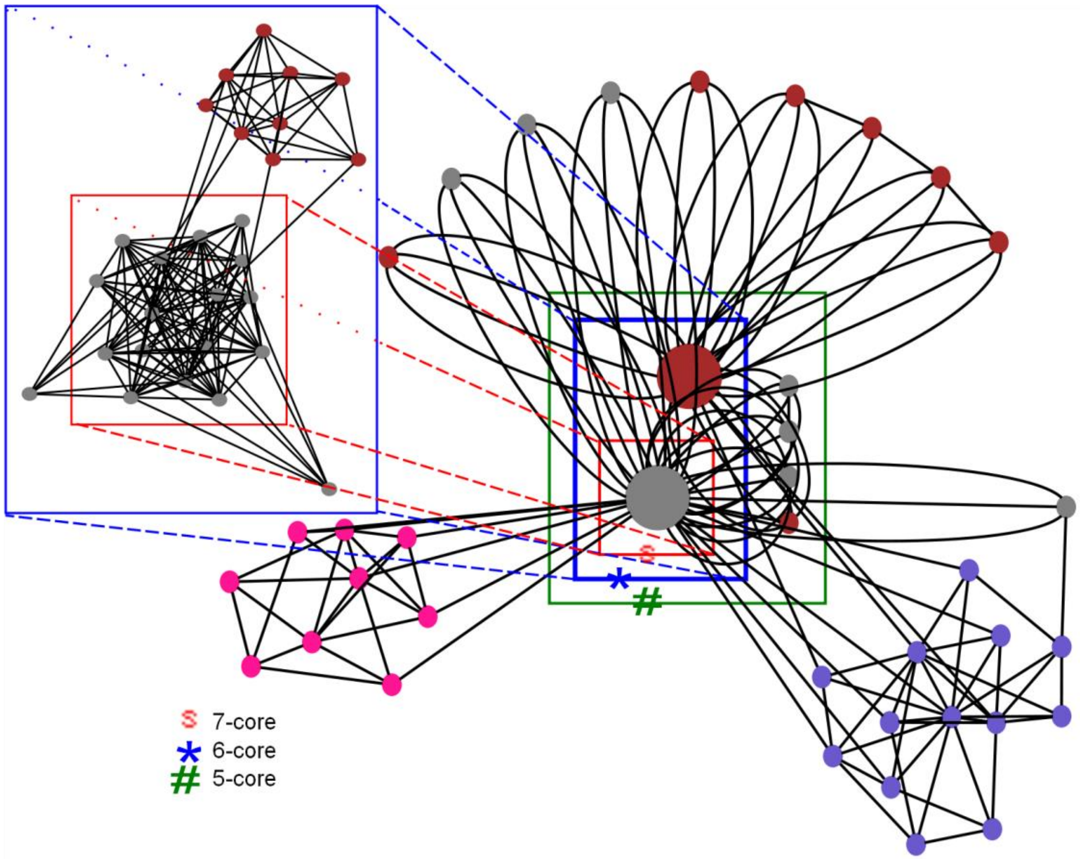
Graph Clustering and Degeneracy

- Assume an "expensive" algorithm **C** (e.g., Spectral Clustering) as a black box
  - It is less expensive to compute in sections of the data separately
- Utilize the vertical partition of k-core decomposition as incremental input to **C**
- Starting at the max(k)-core, for i-core we:
  - Assign with a simple function nodes to existing clusters (from (i+1)-core)
  - Apply **C** to nodes less connected to the existing clusters than sub-graph nodes of {*(i+1)-core*}-{*i-core*}

CoreCluster framework



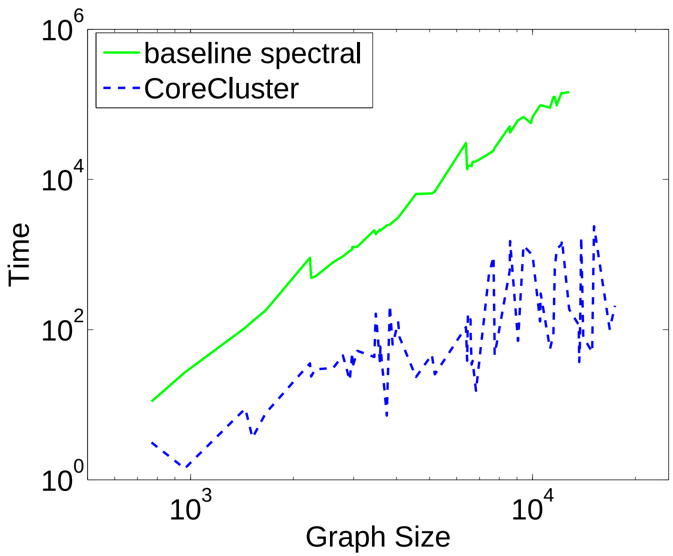
CoreCluster Framework



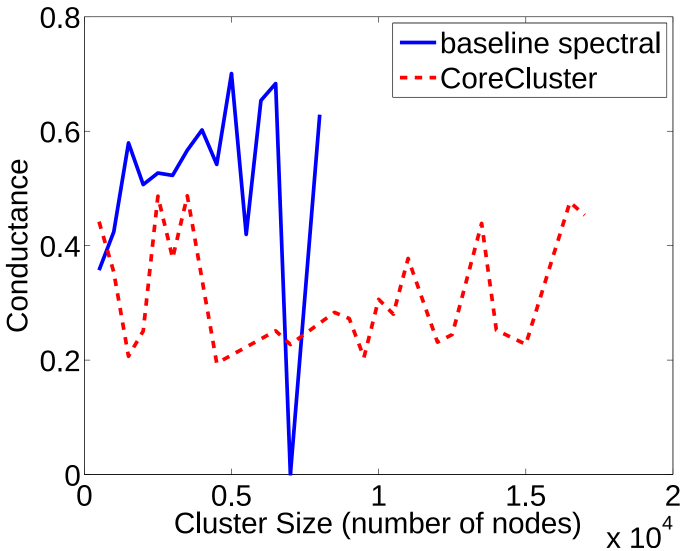
- Core decomposition partitions the graph in a hierarchical nested manner

- We can utilize this structure in graph clustering

Experimental Results (Spectral Clustering)



Execution time



Clustering quality

- Significant improvement in execution time
- Clustering quality is retained

References (alt. methods for community evaluation)

- G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 2005.
- I. Farkas, D. Ábel, G. Palla, and T. Vicsek. Weighted network modules. New J. Phys. 9(180), 2007.
- S. Lehmann, M. Schwartz, and L.K. Hansen. Biclique communities. Phys. Rev. E 78(1), 2008.
- J.M. Kumpula, M. Kivelä, K. Kaski, and J. Saramäki. Sequential algorithm for fast clique percolation. Phys. Rev. E 78, 2008.
- P. Pollner, G. Palla, and T. Vicsek. Parallel clustering with CFinder. Parallel Processing Letters 22, 2012.
- R. Andersen and K.J. Lang. Communities from Seed Sets. In: WWW, 2006.
- R. Andersen, F. Chung, and K.J. Lang. Local Graph Partitioning using PageRank Vectors. In: FOCS, 2006.
- R. Andersen and Y. Peres. Finding Sparse Cuts Locally Using Evolving Sets. In: STOC, 2009.
- D. Gleich and C. Seshadhri. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In: KDD, 2012.
- A.S. Maiya and T.Y. Berger-Wolf. Sampling Community Structure. In: WWW, 2010.

References (alt. methods for community evaluation)

- J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics 6(1), 2009.
- S.L. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. A simple conceptual model for the internet topology. In: GLOBECOM, 2001.
- D. Chakrabarti, Y. Zhan, D. Blandford, C. Faloutsos and G. Blelloch. NetMine: New Mining Tools for Large Graphs. In: SDM Workshop on Link Analysis, Counter-terrorism and Privacy, 2004.
- F.D. Malliaros, V. Megalooikonomou, and C. Faloutsos. Fast robustness estimation in large social graphs: communities and anomaly detection. In: SDM, 2012.
- J. Leskovec, K.J. Lang, and M.W. Mahoney. Empirical comparison of algorithms for network community detection. In: WWW, 2010.
- C. Giatsidis, F. D. Malliaros, D. M. Thilikos, and M. Vazirgiannis. CoreCluster: A Degeneracy Based Graph Clustering Framework. In Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI), 2014.

References (degeneracy)

- C. Giatsidis, D. Thilikos, M. Vazirgiannis,"D-cores: Measuring Collaboration of Directed Graphs Based on Degeneracy", Knowledge and Information Systems Journal, Springer, 2012.
- Christos Giatsidis, Dimitrios M. Thilikos, Michalis Vazirgiannis: D-cores: Measuring Collaboration of Directed Graphs Based on Degeneracy. In: ICDM, 2011.
- Christos Giatsidis, Klaus Berberich, Dimitrios M. Thilikos, Michalis Vazirgiannis: Visual exploration of collaboration networks based ongraph degeneracy. In: KDD, 2012.
- Christos Giatsidis, Dimitrios M. Thilikos, Michalis Vazirgiannis: Evaluating Cooperation in Communities with the k-Core Structure. In: ASONAM, 2011.
- S.B. Seidman. Network Structure and Minimum Degree. Social Networks, 1983.

- An online demo at: **http://www.graphdegeneracy.org/**

New directions for research in the area of graph mining

Μάθημα: Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό

Ενότητα # 5: Community Detection and Evaluation
in Social and Information Networks

Διδάσκων: Μιχάλης Βαζιργιάννης

Τμήμα: Προπτυχιακό Πρόγραμμα Σπουδών “Πληροφορικής”

Open Problems and Future Research Directions (1)

- ***Community detection in directed graphs***
  - A formal and precise definition of the clustering/community detection problem in directed networks (how clusters should look like)
  - In the existing methods on directed networks, there is no a clear way of how the edge directionality should be taken into account
  - Not straightforward generalizations of the methods for undirected graphs
  - **Note:** a single definition/notion of communities should possibly not fit to all needs – highly application-oriented task **[Schaeffer ’07]**
- Extension of existing methods to cover the case of signed graphs

Open Problems and Future Research Directions (2)

- ***Scalability***
  - Distributed spectral clustering
    - Compute Laplacian and eigenvector decomposition in a *distributed* manner
  - Degeneracy for large scale graph clustering
    - Degeneracy identifies the cores of the best clusters
    - The degenerated data are exponentially smaller than the original one so the scheme scales
  - k-core computation O(nm)
    - Can be costly for dense graphs
    - Optimize with divide and conquer + start from high degree nodes

Open Problems and Future Research Directions (3)

- **Clustering Validity for graph clustering**
  - How to decide if the results of graph clustering are valid ?
  - Parameter values and algorithms choice …
  - Reliable benchmark graph dataset **[Lancichinetti and Fortunato ’09]**
  - Experimental and comparative studies should be performed

- **Towards data-driven and application-driven approaches**
  - Study the structure and properties of the graph we are interested in
  - Take into account possible structural observations that may affect the community detection task

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Τέλος Ενότητας # 5

Μάθημα: Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, **Ενότητα # 5:** Community Detection and Evaluation in Social and Information Networks

Διδάσκων: Μιχάλης Βαζιργιάννης**, Τμήμα:** Προπτυχιακό Πρόγραμμα Σπουδών "Πληροφορικής"