Analysis of Annotated Social and Information Networks: Methods and Applications

Miloš Savić

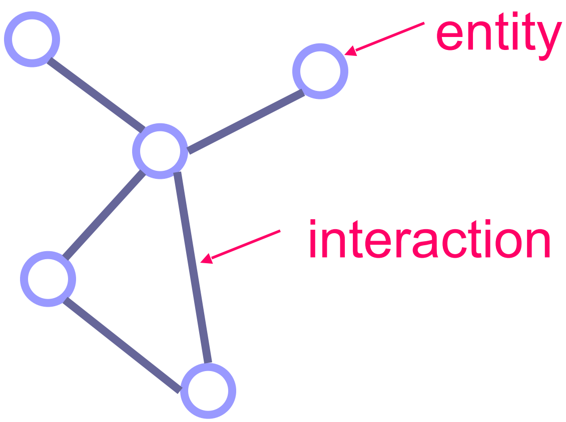
Department of Mathematics and Informatics,
Faculty of Sciences,
University of Novi Sad, Serbia

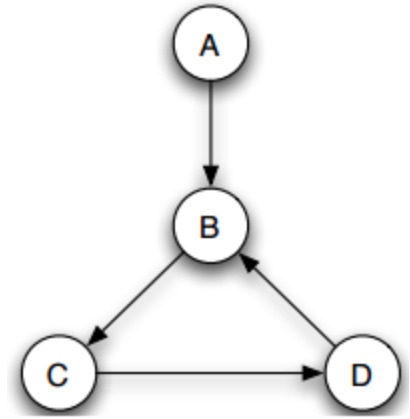
Outline

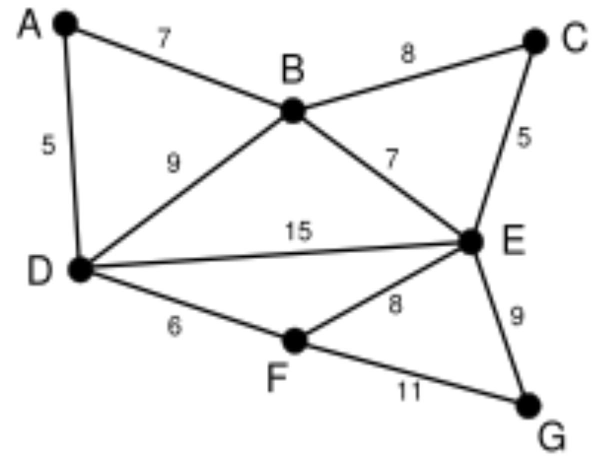
- Introduction
- Fundamentals of complex network analysis
- Methods for annotated networks
- Case study 1 — analysis of enriched co-authorship networks
- Case study 2 — analysis of enriched ontology networks
- Conclusions

Network

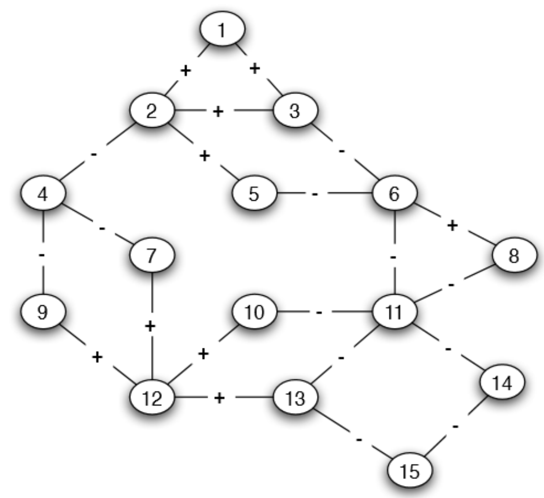
- Network – a graph representing interactions or relations among constituent entities of a complex system







Entities	Interactions	
vertex	edge, arc	math
node	link	computer science
site	bond	physics
actor	tie, relation	sociology



Newman’s classification of complex networks

- **Technological networks**
  - networks representing engineered man-made systems
- **Social networks**
  - Interactions and relationships among social entities
- **Information networks**
  - Connections between data items
- **Biological networks**
  - Networks representing biological systems and processes

Social networks

- **Social network - network-structured data describing interactions or relations among social entities**

- **Social entities**
  - individuals, social groups, institutions, organizations, companies, political parties, nations

- **Social links**
  - opinions on other individuals (signed social networks)
  - transfers of material resources
  - links denoting collaboration, cooperation and coalition
  - links resulting from behavioral interactions
  - links imposed by formal relations within formally organized social groups
  - links on social networking sites
  - …

Information networks

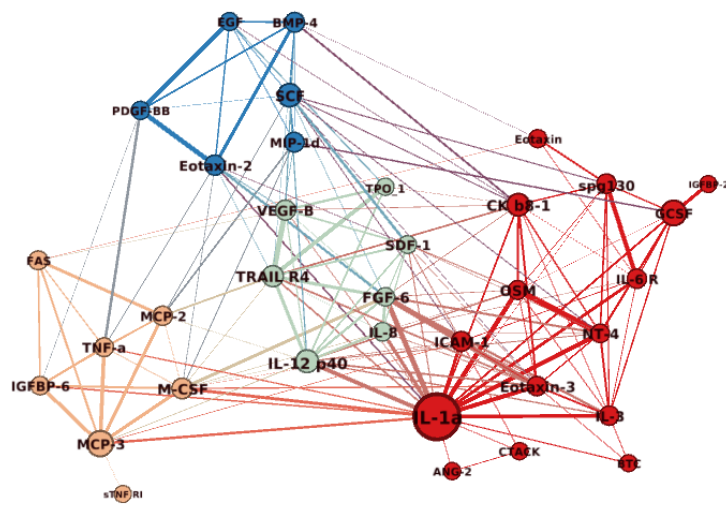
- Networks depicting relations/dependencies between data items
  - **WWW networks**
    - nodes: WWW pages
    - links: hyperlinks (directed links)
  - **Citation networks: references between documents**
    - Scientific papers, patents, legal documents
  - **Linguistic networks**
    - Semantic: semantic relationships (e.g., synonyms or antonyms) between words or concepts
    - Structural: word co-occurrence networks and sentence similarity networks
  - **Recommender networks**
    - Bipartite graphs showing preferences of individuals towards some items
  - **Ontology networks (knowledge graphs)**
    - relationships between ontological entities (concepts, roles, individuals)
    - dependencies between ontology modules of a modular ontology

- **Tabular datasets can be transformed to information networks**
  - **Nodes:** data items or features themselves

- **k-nearest neighbors networks**
  - A → B if B is among the top k nearest data items to A

- **eps-radius networks**
  - A and B connected if distance(A, B) < Eps

- **feature correlation networks**
  - Two features connected if there is a strong correlation between them



Savić et al. A Feature Selection Method Based on Feature Correlation Networks. In P*roc. of MEDI'2017*, pp. 248-261, 2017.

Annotated networks

- Networks whose nodes are augmented with attributes
  - labels/categorical attributes: the value of an attribute restricted to a set of specified categories
  - attributes with numerical values
  - free-text

- **In this tutorial:** networks whose nodes are enriched with both domain-independent metrics used in complex network analysis and domain-dependent metrics
  - **enriched co-authorship networks**
    - metrics quantifying various determinants of research performance
  - **enriched ontology networks**
    - ontology metrics used to evaluate the complexity and design quality of ontologies

Outline

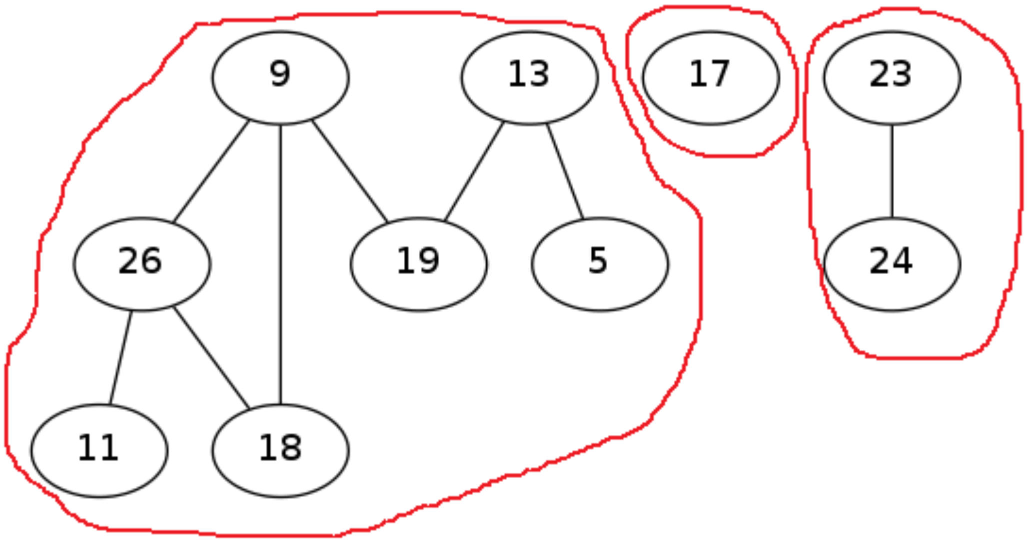
- Introduction
- **Fundamentals of complex network analysis**
- Methods for annotated networks
- Case study 1 — analysis of enriched co-authorship networks
- Case study 2 — analysis of enriched ontology networks
- Conclusions

Complex network analysis

- Quantitative methods for studying the structure and evolution of complex networks
  - Analysis of direct and indirect connectivity of nodes, identification of connectivity trends and patterns
  - Centrality metrics and algorithms — identification of the most important nodes and links in a network
  - Network comprehension — identification of cohesive subgraphs (clusters/communities), analysis of connectivity between and within clusters
  - Identification of evolutionary trends and principles that can explain the evolution of a network at the microscopic, mesoscopic and macroscopic level
  - …

Connected components in undirected networks

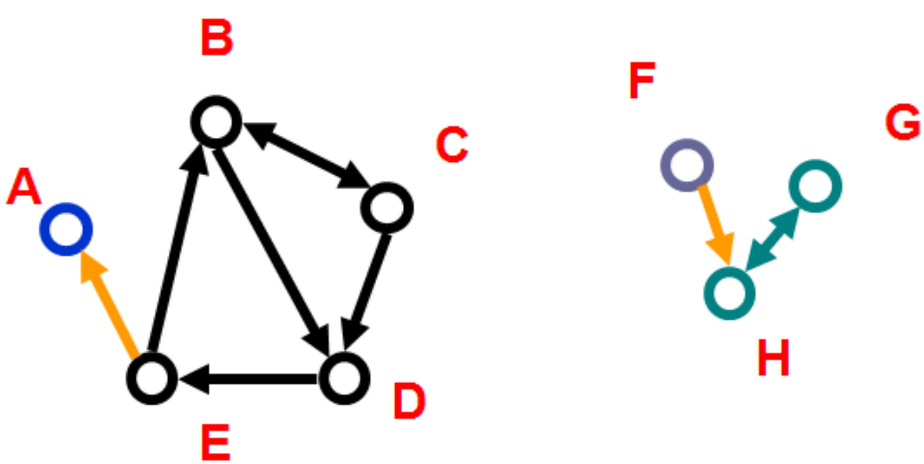
- Connected undirected graph — there is a path between any two nodes
- If a network is not a connected graph then it consists of multiple connected components



- BFS/DFS
- Giant connected component: a component encompassing a vast majority of nodes

Components in directed networks

- **Weakly connected components**
  - connected components in the undirected projection of a directed network
- **Strongly connected components**
  - for every two nodes A and B
    - there is a directed path from A to B, and
    - a directed path from B to A

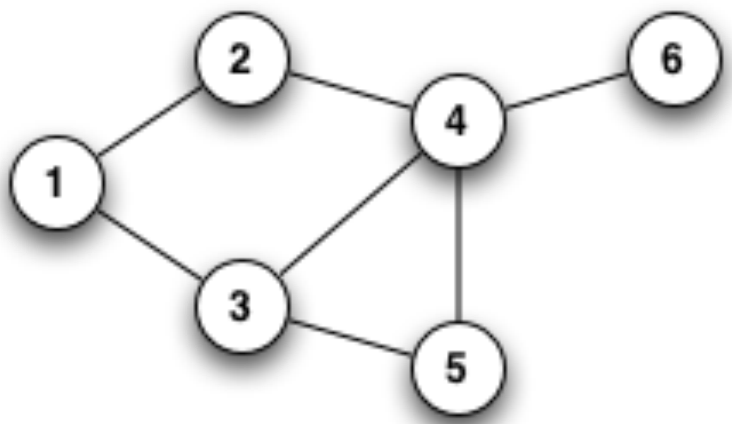


Weakly connected components:
{A, B, C, D, E}
{F, G, H}

Strongly connected components:
{B, C, D, E}
{G, H}

Node degree

- degree(x) = the number of links incident with x
  = the number of x's neighbors
- the most basic metric to assess node importance
  - e.g. in social networks: degree is a metric of social capital
    higher number of contacts → broader possibilities to spread ideas/opinions/interests and influence others
- Directed networks: in-degree and out-degree
- Isolated nodes and hubs



Node	Degree
1	2
2	2
3	3
4	4
5	2
6	1

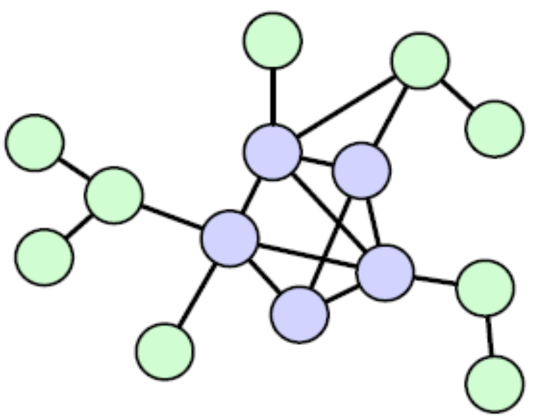
Core-periphery structure

- Assortative networks with localized hubs

- **k-core —** maximal sub-graph S containing nodes whose degree is higher than or equal to k in S

```
void identifyCore(int k) {
      while network contains a node whose degree is < k:
            remove nodes whose degree is < k
      remaining nodes constitute k-core
}
```

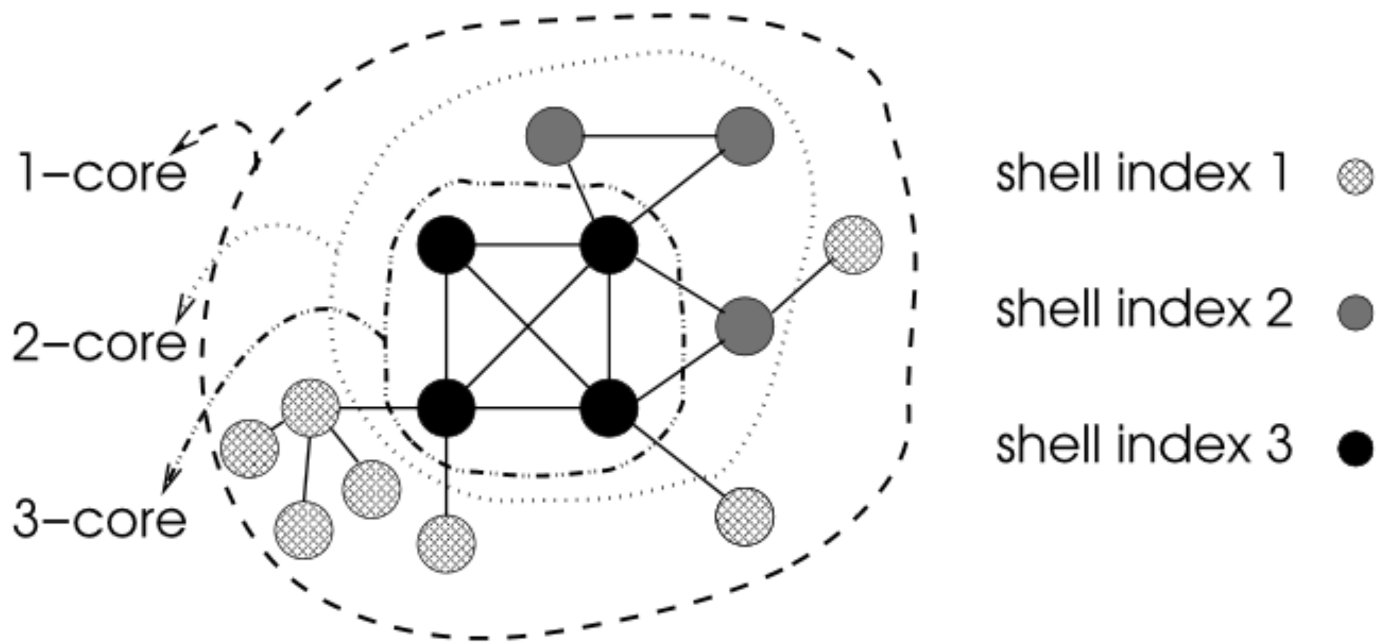
localized hubs:
a k-core for a large k is
a connected graph or
has a giant connected component



core-periphery

k-core decomposition

- k-cores are nested
- shell-index(x) = k — x belongs to k-core, but not to (k+1)-core
- Hubs with
  - high shell-index: hubs connected to other hubs
  - low shell-index: hubs connected to low-degree nodes



Centrality metrics

- Metrics to rank and identify the most important nodes/links in the network

- Fundamental node centrality metrics originate from social network analysis
  - Betweenness centrality
  - Closeness centrality
  - Eigenvector centrality

- Information retrieval
  - centrality metrics for directed graphs inspired by eigenvector centrality
  - Page rank and HITS hub and authority scores

Betweenness centrality

- A node is important if it is located on a large number of shortest paths between other nodes
  - Such node is in a position to control, maintain and influence information flow through the network

Definition 2.38 (Betweenness centrality). The betweenness centrality of a node z in a graph G, denoted by $C_b(z)$, is the extent to which z is located on the shortest paths between two arbitrary nodes different than z:

$$C_b(z) = \sum_{x,y \in V, x \neq y \neq z} \frac{\sigma(x,y,z)}{\sigma(x,y)} \tag{2.10}$$

where $\sigma(x,y)$ is the total number of shortest paths between x and y, and $\sigma(x,y,z)$ is the total number of shortest paths between x and y passing through z.

Closeness centrality

- A node is important if it is in proximity to a large number of other nodes
  - Spreading/diffusion processes: information originating at nodes having a high closeness centrality quickly propagate through the network

Definition 2.41 (Closeness centrality). The closeness centrality of a node z in a graph G, denoted by $C_c(z)$, is inversely proportional to the total distance between z and all other nodes in G:

$$C_c(z) = \frac{1}{\sum_{i \in V \setminus \{z\}} d_{zi}} \tag{2.17}$$

Eigenvector centrality

- Recursively defined centrality: a node is important if it is directly connected to other important nodes

Definition 2.44 (Eigenvector centrality). The eigenvector centrality of a node z in a graph G, denoted by $C_e(z)$, is proportional to the sum of eigenvector centralities of its neighbors:

$$C_e(z) = \frac{1}{\lambda} \sum_{i \in N(z)} C_e(i) \tag{2.21}$$

where λ is a constant and $N(z)$ denotes the set of nodes directly connected to z, i.e. $N(z) = \{w : \{w, z\} \in E\}$.

- EVC can be computed by successive approximations starting from a configuration in which all nodes have equal EVC
- PageRank and HITS hub/authority scores are variants of EVC for directed networks

Node similarity/distance

- **Applications:** community detection (hierarchical agglomerative clustering), link prediction and identification of missing links (in the case of networks extracted from incomplete data)

- The length of the shortest path between two nodes

- Similarity based on random walks: the probability that a random walker reaches *X* from *Y* in *k* random walk steps

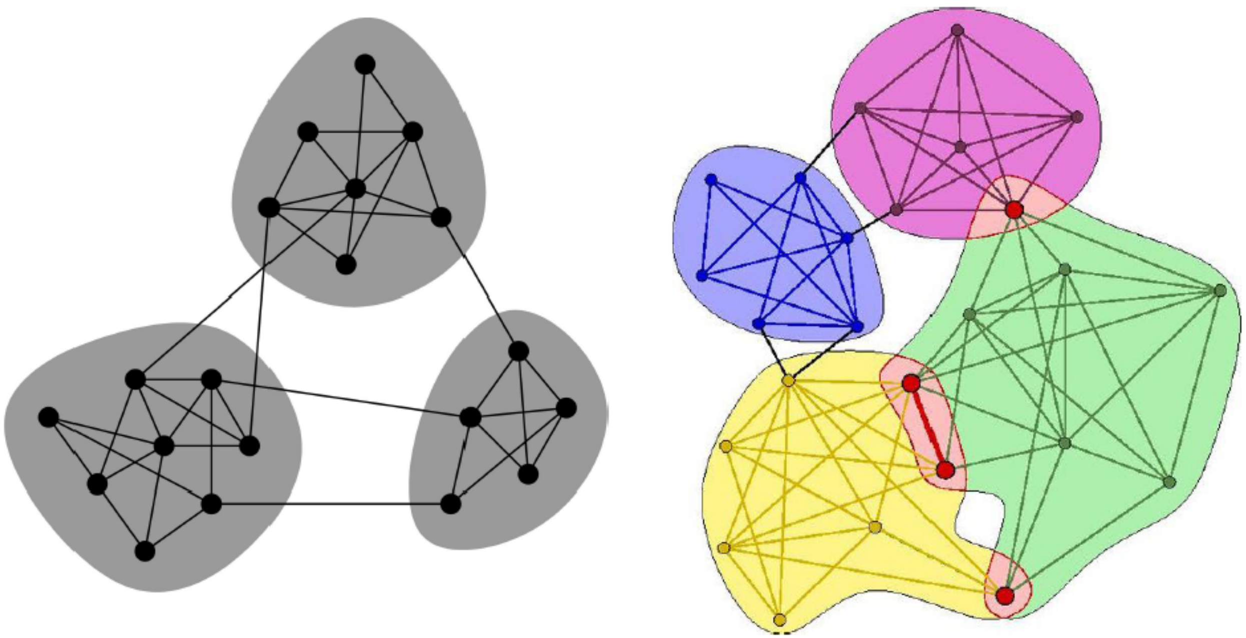
- The number of common neighbors $|\Gamma(x) \cap \Gamma(y)|$

- The Jaccard coefficient $\frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$

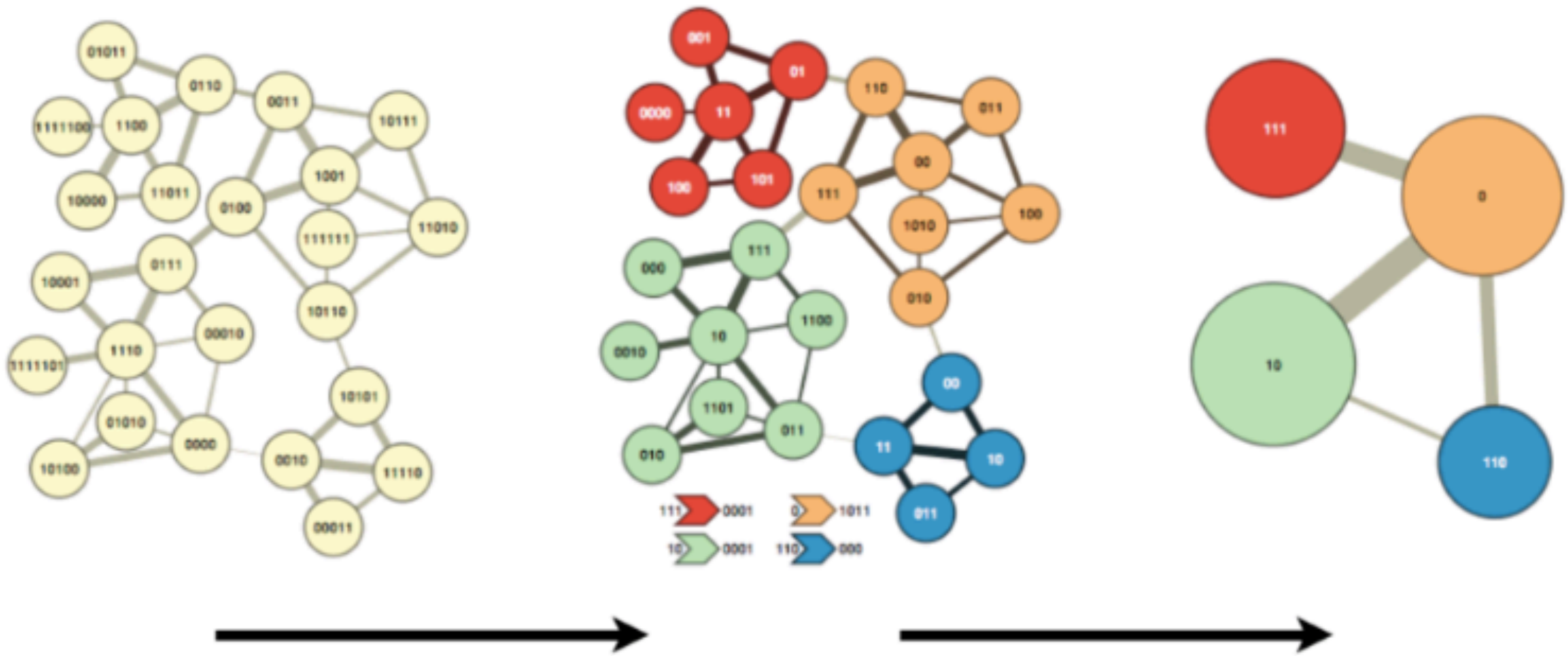
- **Other metrics:** Adamic-Adar, Katz, personalized PageRank, cosine similarity, SimRank (recursively defined similarity)

Community structure

- Community (module, node cluster)
  - a subgraph that is more densely/strongly internally connected than with the rest of the network
- Automatic identification of communities — community detection algorithms
- Overlapping and non-overlapping community partitions



Network comprehension



Find a partition of the network into communities

Coarse-grained description

- Santo Fortunato, 2009, “Community detection in graphs”
  - Agglomerative algorithms
  - Divisive algorithms
    - Repeatedly remove links that are likely to be inter-communitarian links to form the dendrogram
    - Measures indicating inter-communitarian links: edge betweenness centrality, edge clustering coefficient, edge information centrality
  - Modularity-based algorithms
    - heuristics to maximize the modularity measure
      - X — a subgraph in the network
      - Q(X) = the fraction of links in X - the expected fraction of links in X under some null random network model
  - Dynamic algorithms
    - Discovering communities by dynamical processes running on the network (e.g. label propagation)
  - Method-based on statistical inference
    - fitting stochastic block models

Outline

- Introduction
- Fundamentals of complex network analysis
- **Methods for annotated networks**
- Case study 1 — analysis of enriched co-authorship networks
- Case study 2 — analysis of enriched ontology networks
- Conclusion

Analysis of annotated networks

- **Analysis of categorically induced subgraphs**
  - *A* - a categorical node attribute
  - subgraphs induced by nodes having the same value of *A*

- **Are categorically induced subgraphs strong clusters in the network?**
  - **enriched co-authorship networks:** do researchers from the same department form a strongly cohesive research community?
  - **enriched ontology networks:** do ontology modules conform to the "high cohesion" design principle (are concepts from a module strongly related)?
  - **Radicchi et al. notion of clusters in complex networks and graph clustering evaluation (GCE) metrics applied to categorically induced subgraphs in annotated networks**

Radicchi et al. definitions of clusters

- C - a subgraph of a network, x - a node in C
- $k_{int}(x)$ - the number of intra-subgraphs links incident with x, i.e. links connecting x with other nodes in C
  - weighted networks: the total weight of intra-subgraph links
  - directed networks: the number intra-subgraph links emanating from x
- k_{ext}(x) - the number of inter-subgraph links incident with x, i.e. links connecting x with nodes that are not in C

Definition 2.60 (Radicchi strong community). A subgraph C of a graph G is a Radicchi strong community (or community in the strong sense) if

$$(\forall i \in C)\, k_{int}(i) > k_{ext}(i) \tag{2.51}$$

Definition 2.61 (Radicchi weak community). A subgraph C of a graph G is a Radicchi weak community (or community in the weak sense) if

$$\sum_{i \in C} k_{int}(i) > \sum_{i \in C} k_{ext}(i) \tag{2.52}$$

GCE metrics

- C - a subgraph of a network with N nodes
- N_C - the number of nodes in C
- **GCE metrics based on edge-cut**
  - E_C (the size of the edge-cut of C)
    the total number/weight of (out-going) links connecting the nodes in C with nodes that are not in C
  - I_C — the total number/weight of intra-subgraph links in C

  - Conductance(C) = E_C / (E_C + 2I_C) [undirected networks]
  - Conductance(C) = E_C / (E_C + I_C) [directed networks]
  - Expansion(C) = E_C / N_C
  - Cut-ratio(C) = E_C / N(N - N_C) [only for unweighted networks]

  - Lower values of conductance, expansion and cut-ratio indicate more cohesive subgraphs
  - **Conductance(C) < 0.5 → C is a Radicchi weak cluster**

GCE metrics

- C - a subgraph of a network, x - a node in C
- **GCE metrics based on degree-fraction (DF)**
  - $k_{ext}(x)$ - the number/weight of (out-going) inter-subgraph links incident with x
  - $D(x)$ - the (out-) degree/strength of x, $D(x) = k_{int}(x) + k_{ext}(x)$
  - $DF(x) = k_{ext}(x) / D(x)$

  - Maximum-DF(C) = the maximum DF of nodes in C
  - Average-DF(C) = the average DF of nodes in C
  - Flake-DF(C) = the fraction of nodes in C for which $DF(x) < D(x) / 2$ (or, equivalently, $k_{int}(x) > k_{ext}(x)$)

  - Lower values of Maximum-DF and Average-DF and higher values of Flake-DF indicate more cohesive subgraphs
  - **Flake-DF(C) = 1 → C is a Radicchi strong cluster**

Analysis of annotated networks

- **Comparison of categorically induced subgraphs (CISs)**
  - *A* - a categorical node attribute
  - CISs: subgraphs induced by nodes having the same value of *A*
  - *M* - a numerical node attribute

- **Do nodes from a CIS *X* tend to have higher values of *M* compared to nodes from a CIS *Y*?**
  - **enriched co-authorship networks:** do researchers from a department X tend to be more productive/more central in the co-authorship network than researchers from a department Y?
  - **enriched ontology networks:** are concepts from an ontology module X more important than concepts from an ontology module Y?

- **Metric-based comparison test based on the MWU test and probabilities of superiority applied to two categorically induced subgraphs**

Metric-based comparison test

- X and Y - two independent subsets of nodes in a network
- Thr - a probability threshold indicating a strong stochastic dominance

- **Metric-based-comparison-test(X, Y, Thr):**
  - for-each numeric attribute M:
    - M(X) - the set of M values for X
    - M(Y) - the set of M values for Y
    - p = apply the MWU test to M(X) and M(Y)
    - if the null hypothesis rejected ($p < 0.05$):
      - compute probabilities of superiority PS(X) and PS(Y)
        - x = a randomly selected value from M(X)
        - y = a randomly selected value from M(Y)
        - $PS(X) = P(x > y)$, $PS(Y) = P(y > x)$
      - if PS(X) > Thr or PS(Y) > Thr (default Thr = 0.75):
        - **report not only statistically significant differences between X and Y regarding M, but a strong tendency of superiority**

Metric-based comparison test

- Metric-based comparison test can also be applied to two independent sets of nodes determined by some structural criteria, e.g.
  - **highly and lowly coupled nodes**
    - highly coupled nodes: the minimal subset of nodes *C* such that

$$\sum_{x \in C} \text{degree}(x) > \sum_{y \notin C} \text{degree}(y)$$

  - **core and periphery nodes** when the network has a core-periphery structure
    - core nodes: the minimal subset of nodes *C* such that

$$\sum_{x \in C} \text{shell-index}(x) > \sum_{y \notin C} \text{shell-index}(y)$$

  - nodes belonging to non-trivial strongly connected components and nodes not involved in cyclic dependencies in directed networks

Analysis of block models of annotated networks

- *P* - a partition of the set of nodes into *k* node groups

- Block model corresponding to *P*
  - **nodes:** node groups
  - **links:** node groups A and B are connected if there is a node from A connected to a node from B

- Block models of annotated network can be formed in two principal ways:
  - **according to a categorical node attribute**
    - e.g. a departmental collaboration network derived from an intra-institutional co-authorship network

  - **according to a partition obtained after community detection**
    - e.g. a network of research groups obtained after research groups were identified by a community detection algorithm applied to the co-authorship network

Group superiority graphs of annotated networks

- *P* - a partition of the set of nodes into *k* node groups

- The block model corresponding to *P* shows connections among node groups

- **Group superiority graphs (GSG) corresponding to *P* are directed graphs reflecting stochastic dominance among node groups with respect to numerical node attributes**
  - M - a numeric node attribute, A and B two node groups
  - A → B in the GSG of M if nodes in A strongly tend to have higher values of M than nodes in B
  - **GSGs: graphs derived from a block model according to the metric-based comparison test**

Mining attachment preferences in annotated networks

- **To which nodes new nodes connect when joining a network?**

- N — an annotated network with k numeric attributes $M_1, M_2, \ldots, M_k$
- N_a and N_b — two successive evolutionary snapshots of N
- Transition from to N_a to N_b
  - New nodes — nodes in N_b not present in N_a
  - Nodes in N_a can be divided into two categories
    - Preferential nodes — nodes to which new nodes attached
    - Non-preferential nodes — nodes that are not preferential

- **Attachment preferences in the evolutionary transition from N_a to N_b can be revealed by the metric-based comparison test**
  - e.g. {(M_3, PREF), (M_8, NON-PREF), (M_{12}, PREF)}

- An algorithm for mining frequent itemsets (e.g. Apriori) applied to the set of attachment preferences of all evolutionary transitions

Outline

- Introduction
- Fundamentals of complex network analysis
- Methods for annotated networks
- **Case study 1 — analysis of enriched co-authorship networks**
- Case study 2 — analysis of enriched ontology networks
- Conclusions

Research collaboration

- **Collaboration:** key social feature of modern science
- **Science from a social perspective:** complex self-organizing social system
- **Katz:** “scientific collaboration is a social process and probably there are as many reasons for researchers to collaborate as there are reasons for people to communicate”
- **Research collaboration can be studied at various levels:**
  - intra-institutional, inter-institutional, national, international, disciplinary, inter-disciplinary
- **Major research questions:**
  - how research collaboration is structured?
  - how the structure of research collaboration evolves?
  - how research collaboration is related to research productivity and impact of multi-authored publications?

Research collaboration

- Research collaboration may manifest in various formal and informal forms

- Co-authorship — the most visible and well-documented manifestation of scientific collaboration
  - availability of massive bibliographic databases

- Co-authorship networks — social networks encompassing researchers
  - Nodes — researchers
  - A and B are connected if A and B co-authored at least one publication (with or without other co-authors)
  - Link weights — the strength of research collaboration

Link weighting schemes

- Straight scheme
  - w(x, y) = the number of joint publications of x and y

- Salton's scheme — a normalized variant of the straight scheme

$$w = \frac{h_{x,y}}{\sqrt{h_x \cdot h_y}}$$

h(x) — the number of publications (co-)authored by x
h(y) — the number of publications (co-)authored by y
h(x, y) — the number of joint publications of x and y

- Newman's scheme
  - More authors a paper has less weight should be added to the total strength of research collaboration

$$w = \sum_{k \in J} \frac{1}{n_k - 1}$$

J — the set of joint publications of x and y
n(k) — the number of authors of publication k

- ## Case study: the FS-UNS co-authorship network
  - The network reflecting intra-institutional research collaboration at FS-UNS (423 FS-UNS researchers from 5 departments)
  - The network extracted from bibliographic records contained in the institutional CRIS-UNS system
    - No name disambiguation problems
    - Categorization of publications by the rule book prescribed by the Serbian Ministry of Science

      → Serbian research competency index metric
  - The Newman schema used to assign link weights
  - Nodes enriched with metrics quantifying different determinants of research performance

Department	Abbrv.
Department of Biology and Ecology	DBE
Department of Physics	DP
Department of Geography, Tourism and Hotel Management	DG
Department of Chemistry, Biochemistry and Environmental Protection	DC
Department of Mathematics and Informatics	DMI

Metric	Abbreviation	Category
Productivity, normal count	PRON	Productivity
Productivity, fractional count	PROF	Productivity
Productivity, straight count	PROS	Productivity
Serbian Research Competency Index	SRCI	Productivity
The total number of co-authors	COLL	Collaboration
The number of FS-UNS co-authors	LCOLL	Collaboration
The number of external co-authors	ECOLL	Collaboration
The strength of research collaboration with all co-authors	WCOLL	Collaboration
The strength of research collaboration with FS-UNS co-authors	WLCOLL	Collaboration
The strength of research collaboration with external co-authors	WECOLL	Collaboration
Clustering coefficient	CC	Collaboration
The degree of intra-group collaboration	IntraDEG	Collaboration
The degree of inter-group collaboration	InterDEG	Collaboration
The strength of intra-group collaboration	WIntraDEG	Collaboration
The strength of inter-group collaboration	WInterDEG	Collaboration
Betweenness centrality	BET	Importance
Weighted betweenness centrality	WBET	Importance
Closeness centrality	CLO	Importance
Weighted closeness centrality	WCLO	Importance
Eigenvector centrality	EVC	Importance

Cohesiveness of research departments

- All FS-UNS departments are Radicchi weak and close to Radicchi strong clusters in the network
  - Intra-department collaborations are stronger than inter-department collaborations for a large majority of researchers but not for all of them
- The strongest intra-department collaborations: DP and DC
- The weakest intra-department collaborations: DMI
- The most closed department: DG (the highest internal density, the lowest conductance)

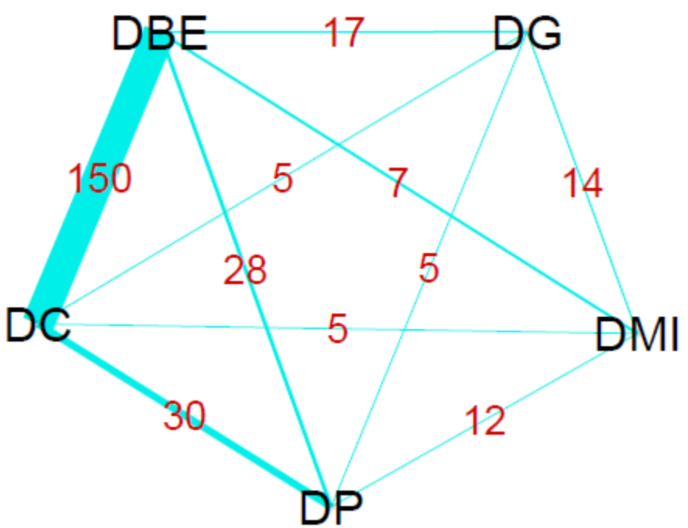
Metric	DBE	DP	DC	DMI	DG
The number of researchers	118	57	95	87	66
The number of non-trivial components	1	1	1	1	1
The number of isolated nodes	3	3	0	7	6
The number of intra-department links	660	240	617	197	560
The number of inter-department links	412	174	411	71	96
Internal density	0.096	0.15	0.14	0.05	0.26
Total weight of intra-department links	8073	5636	9261	1532	2513
Total weight of inter-department links	1607	683	1825	195	160
Average internal degree	11.19	8.42	12.99	4.53	16.97
Average internal weighted degree	136.83	197.76	194.97	35.22	76.15
Weighted conductance	0.17	0.11	0.16	0.11	0.06
Weighted Flake degree fraction	0.97	0.93	0.98	0.95	0.95

Inter-department collaborations

- Researchers involved in inter-department collaborations are drastically more productive, collaborative and institutionally important

Node metric	Avg(G_1)	Avg(G_2)	U	p	PS_1	PS_2
SRCI	160.378	58.6939	11178.5	1.08E-18*	0.7482	0.2507
PRON	104.9031	32.9031	10333	2.06E-21*	0.764	0.2285
PROS	29.2555	13	13781	1.40E-11*	0.6764	0.2959
PROF	27.9682	12.3087	13477.5	2.69E-12*	0.697	0.3029
LCOLL	18.7225	7.4592	7486.5	4.92E-32*	0.82	0.1566
ECOLL	51.0088	13.4745	8411.5	2.52E-28*	0.8038	0.1819
COLL	69.7313	20.9337	7360	1.62E-32*	0.8304	0.1612
BET	769.6687	98.0929	7775	5.09E-31*	0.8166	0.1661

- The departmental collaboration network of FS-UNS is a clique,
but the strengths of inter-department collaborations are highly unbalanced
(a lot of space to improve inter-department collaborations)



Metric-based comparison of departments

- Kruskal-Wallis ANOVA: statistically significant differences (SSD) present regarding SRCI and PRON, but absent regarding PROS and PROF
  - **SRCI and PRON – biased measures of productivity**
- SSD in both local and external collaboration
- No SSD regarding institutional importance

Metric	DBE	DP	DC	DMI	DG	χ^2	p-value
SRCI	91.86	174.58	151.84	94.96	67.17	26.01	3.15E-05*
PRON	74.77	98.37	90.54	44.75	50.58	22.68	1.47E-04*
PROS	19.74	25.68	23.48	21.3	19.88	7.85	0.097
PROF	19.17	23.48	20.69	22.2	19.15	6.38	0.172
LCOLL	14.68	11.47	17.32	5.34	18.42	99.11	1.52E-20*
ECOLL	39.17	41.65	43.59	12.36	30.42	49.11	5.54E-10*
COLL	53.85	53.12	60.91	17.7	48.85	69.71	2.61E-14*
BET	514.21	464.53	362.39	553.4	366.87	3.24	0.51811

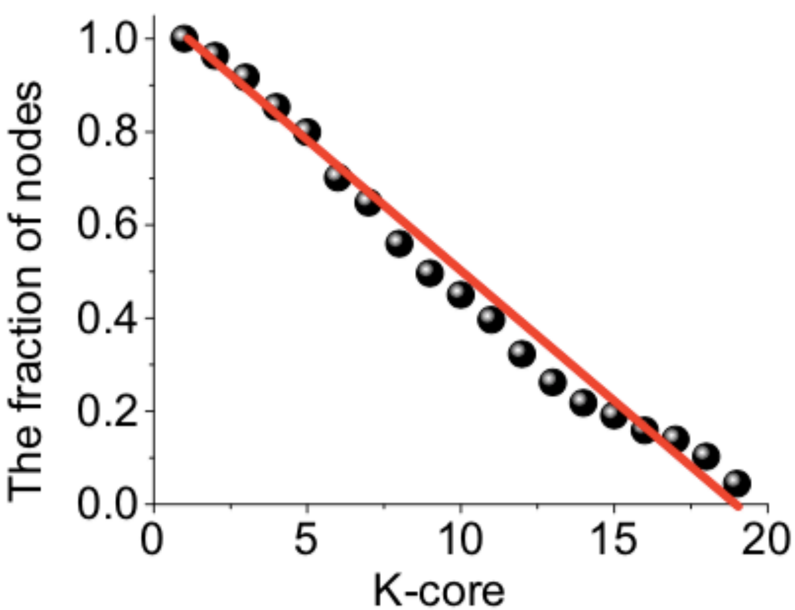
Post-hoc pairwise comparison

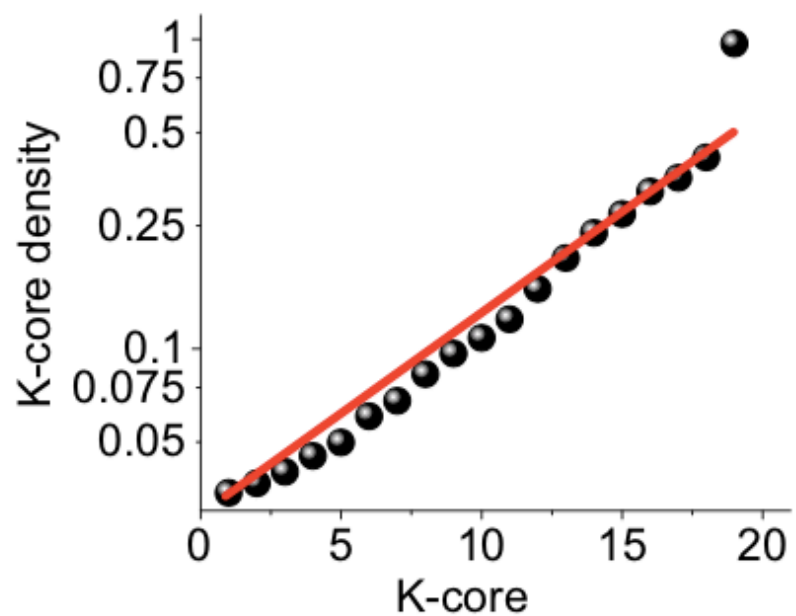
- DP and DC: superior regarding SRCI and PRON
- DMI: the lowest degree of both local and external research collaboration
- DC and DG: active stimulation of intra-institutional collaboration

Metric	Department 1	Department 2	U	p-value	PS_1	PS_2
SRCI	DG	DP	1280	0.0023	0.34	0.66
	DBE	DP	2516.5	0.0071	0.37	0.62
	DP	DMI	1825.5	0.0076	0.63	0.37
	DG	DC	1956	0.0001	0.31	0.69
	DBE	DC	4046.5	0.0005	0.36	0.64
	DMI	DC	2884	0.0004	0.35	0.65
PRON	DP	DMI	1824.5	0.0075	0.63	0.36
	DG	DC	2236.5	0.002	0.35	0.64
	DBE	DC	4345	0.0048	0.38	0.61
	DMI	DC	2481	$< 10^{-4}$	0.29	0.69
LCOLL	DG	DBE	2913	0.0046	0.62	0.36
	DG	DP	1094.5	0.0001	0.70	0.28
	DG	DMI	851	$< 10^{-4}$	0.84	0.14
	DBE	DMI	2297.5	$< 10^{-4}$	0.75	0.20
	DP	DMI	1150.5	$< 10^{-4}$	0.75	0.21
	DBE	DC	4521	0.0153	0.39	0.58
	DP	DC	1888.5	0.0018	0.33	0.63
	DMI	DC	1073	$< 10^{-4}$	0.12	0.86
ECOLL	DG	DMI	1576	$< 10^{-4}$	0.71	0.26
	DBE	DMI	2906	$< 10^{-4}$	0.70	0.27
	DP	DMI	1319	$< 10^{-4}$	0.72	0.25
	DMI	DC	1879	$< 10^{-4}$	0.22	0.76

K-core decomposition

- The FS-UNS co-authorship network has a strong and balanced nested core-periphery structure
  - 19 cores, all of them being connected subgraphs in the network
  - the density of cores increases exponentially
  - the fraction of nodes in k-cores decreases linearly with k
  - **Core researchers: shell-index >= 12 (32% of the total number)**





Core VS Peripheral Researchers

- Core researchers are drastically more productive, collaborative and institutionally important than peripheral researchers.
- Core researchers have more significant brokerage role within their ego-networks

Metric	Avg(C)	Avg(P)	U	p	NHA	PS_1	PS_2
PRON	124.2576	49.6354	7954	$< 10^{-4}$	no	0.78	0.22
PROF	31.7894	16.1697	10003	$< 10^{-4}$	no	0.73	0.27
PROS	32.4924	17.3430	10647	$< 10^{-4}$	no	0.70	0.28
SRCI	172.8083	89.7819	9684	$< 10^{-4}$	no	0.74	0.26
COLL	88.2273	29.8051	4653	$< 10^{-4}$	no	0.87	0.13
LCOLL	26.4697	8.0072	784.5	$< 10^{-4}$	no	0.98	0.02
ECOLL	61.7576	21.7978	7191	$< 10^{-4}$	no	0.80	0.19
WCOLL	120.2045	46.2960	7723	$< 10^{-4}$	no	0.79	0.21
WLCOLL	66.9061	22.8109	6641	$< 10^{-4}$	no	0.82	0.18
WECOLL	53.2984	23.4851	10089.5	$< 10^{-4}$	no	0.72	0.28
BET	813.9461	312.2748	8040	$< 10^{-4}$	no	0.78	0.22
CLO	0.3457	0.2897	4622	$< 10^{-4}$	no	0.87	0.13
EVC	0.0046	0.0014	849	$< 10^{-4}$	no	0.98	0.02
WBET	563.0291	221.8833	12644.5	$< 10^{-4}$	no	0.54	0.23
WCLO	0.6649	0.5056	7099	$< 10^{-4}$	no	0.81	0.19
CC	0.4659	0.5829	13654	$< 10^{-4}$	no	0.37	0.63

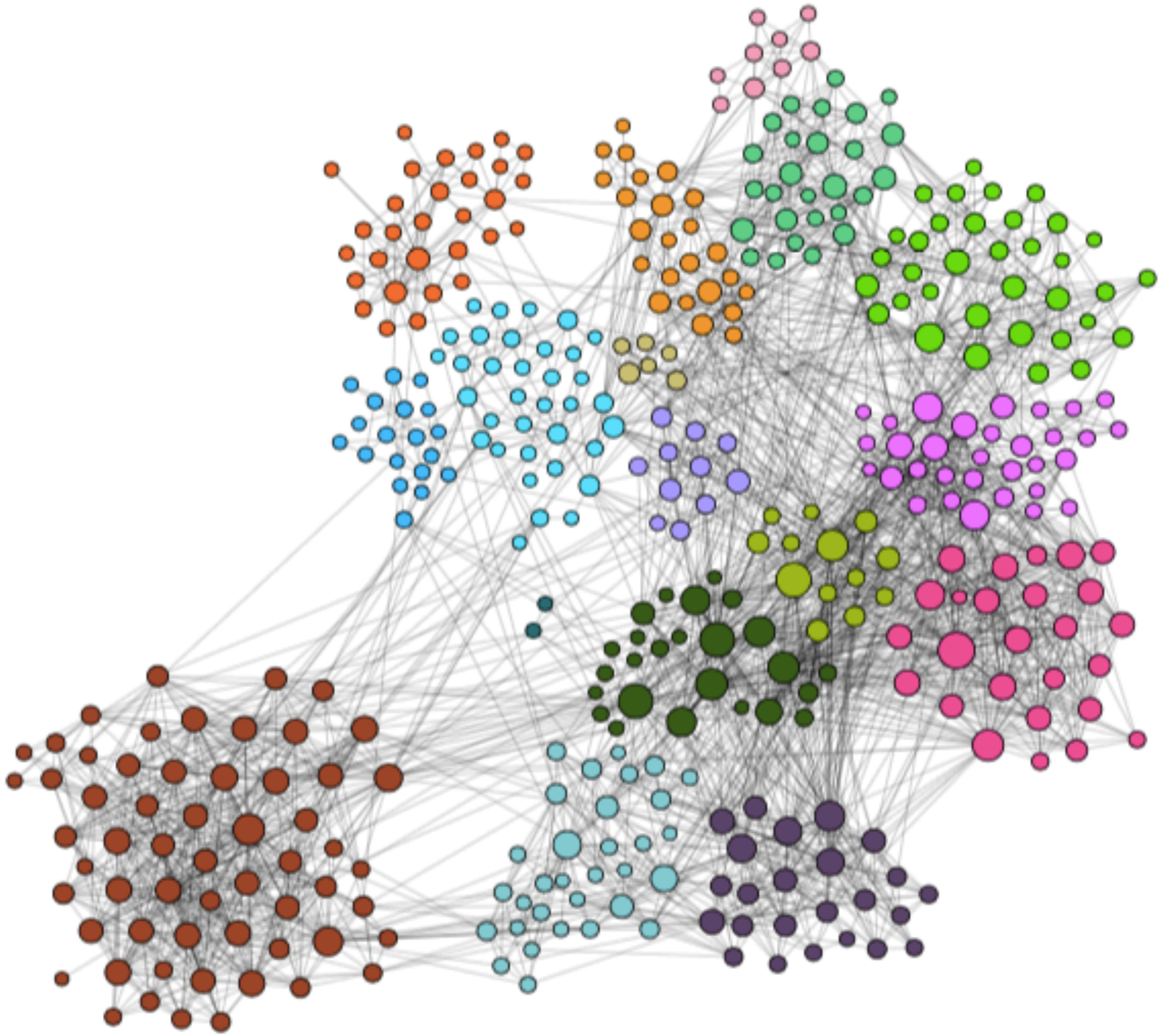
Identification or research groups

Algorithm	Reference	Q	NC	w^{intra}	w^{inter}	r
GMO	[5]	0.8371	18	6919.45	655.66	0.0947
IM	[21]	0.8141	41	6618.53	956.58	0.1445
LV	[3]	0.8466	17	6920.37	654.74	0.0946
WT	[18]	0.8207	37	6873.07	702.04	0.1021
EB	[8]	0.5486	13	5248.49	2326.63	0.4433
SOM	[16]	0.6022	27	6466.84	1108.28	0.1714

- the best performing algorithm: LV (Louvain)
  - the highest modularity, the lowest ratio of w(inter) and w(intra)
- agglomerative clustering techniques better than divisive

$$\text{LV} \succ \text{GMO} \succ \text{WT} \succ \text{IM} \succ \text{SOM} \succ \text{EB}$$

Fig. 8.3: The visualization of the FS-UNS co-authorship network after community detection by the Louvain algorithm. Nodes in the same color belong to the same community. The size of a node is proportional to its degree centrality.

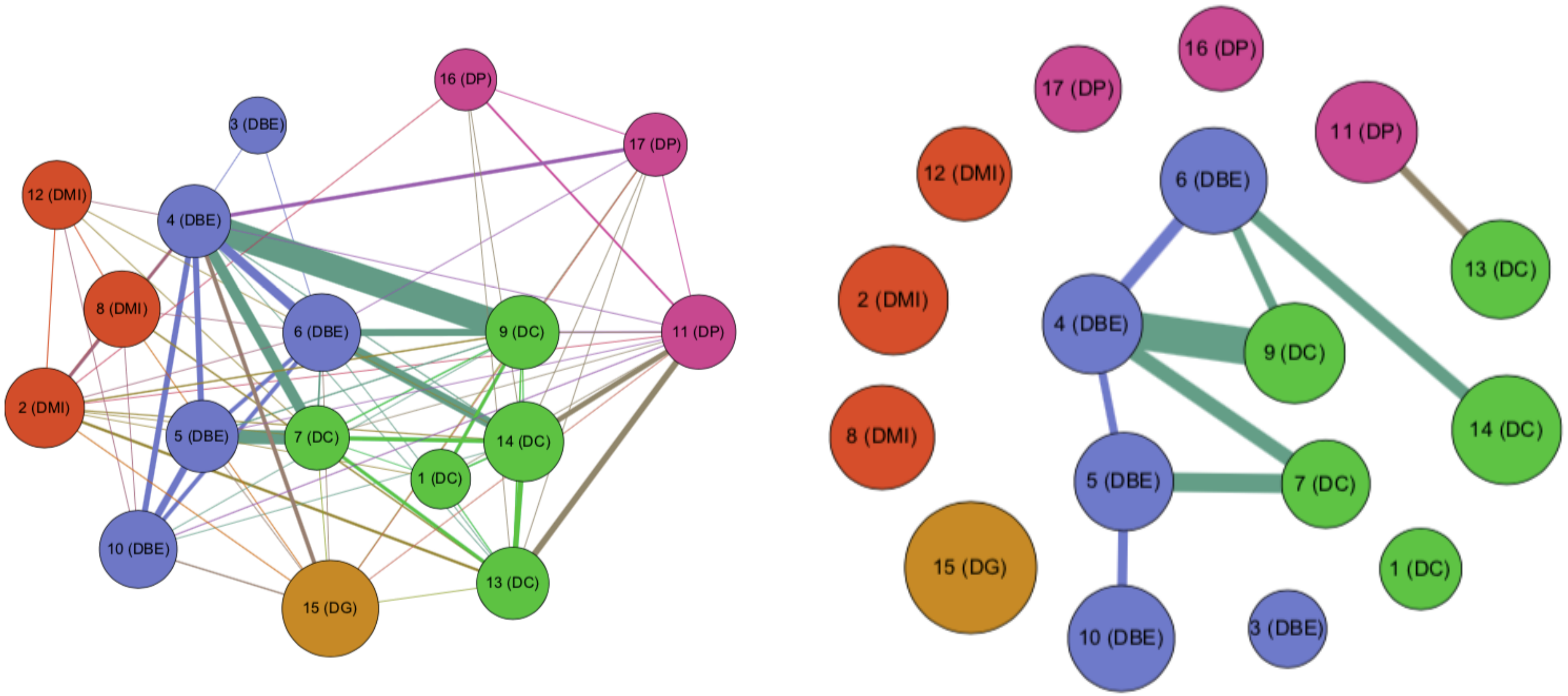


Research groups identified by Louvain

ID	Size	DBE	DP	DC	DMI	DG	DD	w^{intra}	w^{inter}	RS
C1	6	0	0	6	0	0	DC	179.33	23.58	no
C2	35	0	1	1	32	1	DMI	670.30	48.75	no
C3	2	2	0	0	0	0	DBE	82.57	2.70	yes
C4	25	23	0	1	0	1	DBE	927.73	260.12	no
C5	24	24	0	0	0	0	DBE	1230.69	117.35	no
C6	32	32	0	0	0	0	DBE	797.62	136.49	no
C7	13	0	0	13	0	0	DC	712.20	131.51	yes
C8	30	0	1	0	29	0	DMI	843.72	16.83	yes
C9	26	1	0	25	0	0	DC	1848.67	163.32	yes
C10	32	32	0	0	0	0	DBE	927.99	73.69	yes
C11	27	0	21	5	0	1	DP	761.24	68.34	yes
C12	19	0	0	0	19	0	DMI	299.17	10.23	yes
C13	24	0	10	13	1	0	DC	537.04	82.45	yes
C14	35	2	2	31	0	0	DC	948.64	107.32	yes
C15	59	0	0	0	0	59	DG	1642.18	30.58	yes
C16	9	0	9	0	0	0	DP	321.13	11.88	yes
C17	11	0	11	0	0	0	DP	1110.53	24.35	yes

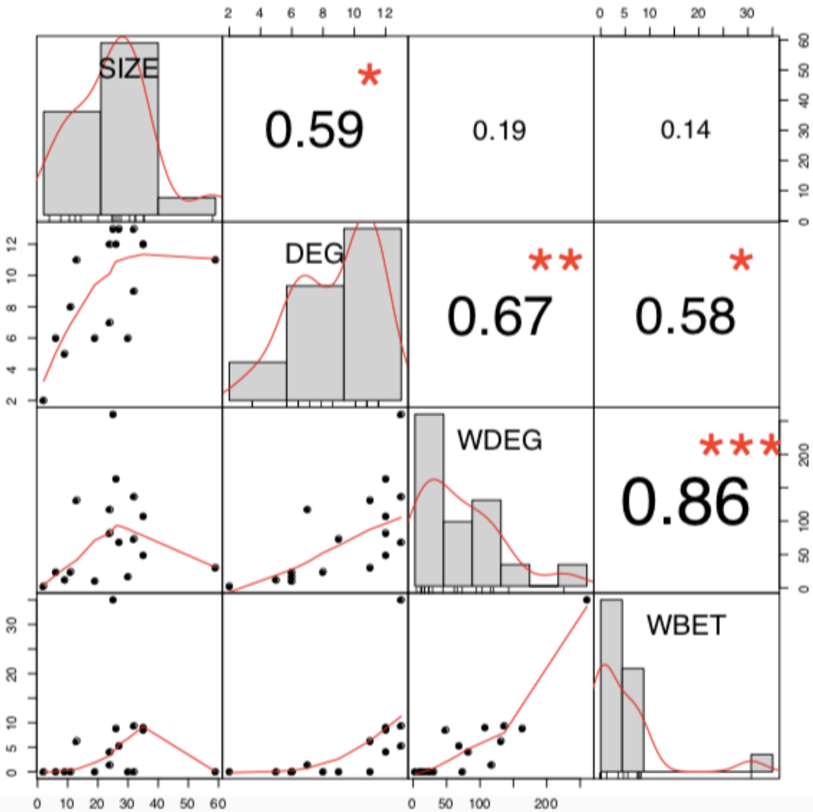
Collaborations among research groups

- The block model formed according to the partition of nodes obtained by the Louvain algorithm
  - #nodes = 17 (research groups)
  - #links = 79 (collaborations between research groups), 9 strong links



- **Expected:** groups that have strong collaborations tend to be more institutionally important
- **The importance of a research group and the strength of inter-group research collaboration are independent of group size**

Fig. 8.6: The Spearman correlation matrix of the size, degree (DEG), weighted degree (WDEG) and weighted betweenness centrality (WBET) of nodes in the collaboration network of FS-UNS research groups.



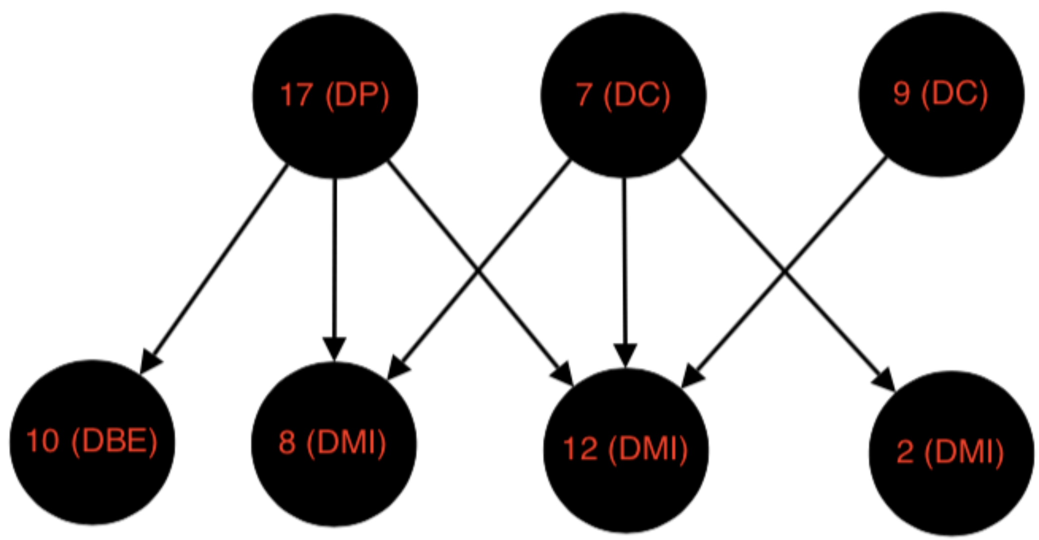
- **Researchers involved in inter-group research collaborations are significantly more productive, collaborative and institutionally important**

Metric	Avg(G_1)	Avg(G_2)	U	p	NHA	PS_1	PS_2
PRON	98.6367	26.8662	8221.5	$< 10^{-4}$	no	0.78	0.21
PROF	26.5373	11.1954	11043	$< 10^{-4}$	no	0.71	0.29
PROS	27.8801	11.6127	11325	$< 10^{-4}$	no	0.69	0.28
SRCI	151.3667	51.1648	9037.5	$< 10^{-4}$	no	0.76	0.24
COLL	65.1873	17.5845	5301	$< 10^{-4}$	no	0.86	0.14
LCOLL	17.4569	7.4014	6744.5	$< 10^{-4}$	no	0.81	0.16
ECOLL	47.7303	10.1831	5932	$< 10^{-4}$	no	0.84	0.15
WCOLL	95.1948	23.0563	7651.5	$< 10^{-4}$	no	0.80	0.20
WLCOLL	48.7996	14.9347	8520.5	$< 10^{-4}$	no	0.78	0.22
WECOLL	46.3951	8.1216	8065	$< 10^{-4}$	no	0.79	0.21
IntraDEG	10.3408	7.4014	12475.5	$< 10^{-4}$	no	0.64	0.30
InterDEG	7.1161	0.0000	0	$< 10^{-4}$	no	1.00	0.00
WIntraDEG	43.8952	14.9347	9557	$< 10^{-4}$	no	0.75	0.25
WInterDEG	4.9045	0.0000	0	$< 10^{-4}$	no	1.00	0.00
BET	687.4335	73.2130	5683	$< 10^{-4}$	no	0.84	0.14
CLO	0.3291	0.2676	3142	$< 10^{-4}$	no	0.92	0.08
EVC	0.0031	0.0013	6730	$< 10^{-4}$	no	0.82	0.18
WBET	457.4900	95.9977	12196.5	$< 10^{-4}$	no	0.51	0.16
WCLO	0.6060	0.4651	8647.5	$< 10^{-4}$	no	0.77	0.23
CC	0.4866	0.6553	11824	$< 10^{-4}$	no	0.30	0.68

- Comparison of research groups by analyzing group superiority graphs corresponding to productivity and collaboration metrics
- **PRON and SRCI — biased measures of research productivity**

Metric	Nodes	Links	Superior groups	Inferior groups	Bipartite structure
PRON	7	7	3	4	yes
PROF	0	0	/	/	/
PROS	0	0	/	/	/
SRCI	11	10	2	9	yes
COLL	13	24	9	4	yes
WCOLL	10	12	4	6	yes

Fig. 8.8: The group superiority graph corresponding to the PRON research productivity metric.



Mining attachment preferences

- New FS-UNS researchers tend to attach to highly productive FS-UNS researchers that have established a strong collaboration with their previous co-authors

Itemset size	Itemset	Support
1	{(PRON, pref)}	0.625
	{(SRCI, pref)}	0.667
	{(COLL, pref)}	0.667
	{(LCOLL, pref)}	0.583
	{(WCOLL, pref)}	0.708
	{(IntraDEG, pref)}	0.583
	{(EVC, pref)}	0.625
2	{(PRON, pref), (SRCI, pref)}	0.625
	{(PRON, pref), (WCOLL, pref)}	0.625
	{(SRCI, pref), (WCOLL, pref)}	0.625
	{(COLL, pref), (LCOLL, pref)}	0.583
	{(COLL, pref), (WCOLL, pref)}	0.625
	{(LCOLL, pref), (WCOLL, pref)}	0.583
	{(WCOLL, pref), (IntraDEG, pref)}	0.583
	{(WCOLL, pref), (EVC, pref)}	0.583
3	{(PRON, pref), (SRCI, pref), (WCOLL, pref)}	0.625
	{(COLL, pref), (LCOLL, pref), (WCOLL, pref)}	0.583

Outline

- Introduction
- Fundamentals of complex network analysis
- Methods for annotated networks
- Case study 1 — analysis of enriched co-authorship networks
- **Case study 2 — analysis of enriched ontology networks**
- Conclusions

- **Ontology - a formal specification of shared and reusable knowledge**
  - Description of concepts (classes) and roles (relationships) in a knowledge domain through a set of axioms in a description logic
  - Backbone of the Semantic Web, specified in OWL
- Monolithic and modular ontology designs
  - **monolithic** — all captured concepts, roles, axioms and assertions gathered together in one (large) OWL file
  - **modular** — an ontology that consists of multiple ontology modules, OWL import feature
- Ontology networks — directed graphs showing dependencies between ontological entities
  - Ontology module networks (nodes: ontology modules, links: import relations between modules)
  - Ontology class networks (nodes: classes, links: relations between classes)
  - Ontology subsumption network (nodes: classes, links: subsumption relations between classes)

Modular design principles

- “Low coupling, high cohesion”
  - an ontology module should be loosely coupled to other ontology modules
    → a low average node degree and the absence of hubs in ontology module networks

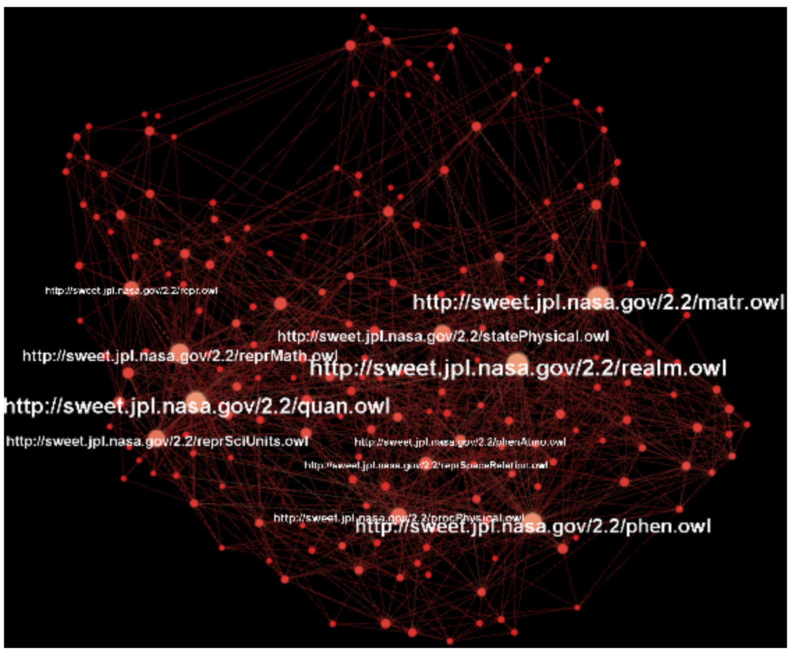
  - concepts in an ontology module should be strongly coupled
    → concepts from the same module form highly cohesive subgraphs (strong clusters) in the ontology class network
    → **GCE metrics as metrics of ontology module cohesion**
    - classification of modules as Radicchi strong, Radicchi weak and poorly cohesive ontology modules
    - Existing ontology cohesion metrics estimate the cohesiveness of ontology modules in isolation (dependencies to external classes are ignored)
    - **GCE metrics rely on external class dependencies taking into account also the principle of “low coupling”**

Modular design principles

- "Avoid cyclic dependencies"
  - Ontology modules belonging to a strongly connected component of the ontology module network are mutually (directly or indirectly) dependent
  - Large cyclic dependencies negatively impact the following quality attributes:
    - understandability
    - reusability
    - maintainability
  - Analysis of strongly connected components in enriched ontology modules networks in order to reveal characteristics of modules involved in cyclic dependencies
    - **Metric-based comparison test**

Case study: enriched ontology module and class networks of SWEET (Semantic Web for Earth and Environmental Terminology)

Network	Abbr.	The number of nodes	The number of links
Ontology module network	OMN	203	1138
Ontology class network	OCN	6374	8483
Ontology subsumption network	OSN	6003	6202



- Module network
  - metrics of internal complexity
  - adopted software metrics
  - centrality metrics
  - Orme et al. coupling metrics
  - Tartir et al. diversity metrics
  - GCE metrics
- Class networks
  - adopted software metrics
  - centrality metrics

Connected component analysis

- **SWEET exhibits a high degree of modular and conceptual cohesion**
  - The SWEET OMN is a weakly connected digraph
    no isolated modules or small independent clusters of modules
  - The SWEET OCN is a weakly connected digraph, a giant connected component in the taxonomy of concept
- All three examined networks are small-world networks

Table 4.3: Weakly connected components in the SWEET ontology networks. #WCC – the number of weakly connected components, LWCCN – the fraction of nodes in the largest WCC, LWCCL – the fraction of links in the largest WCC, SW – the small-world coefficient, SW-rnd – the small-world coefficient of a comparable random graph, CC – the clustering coefficient, CC-rnd – the clustering coefficient of a comparable random graph, *A* – the Newman assortativity index.

Network	#WCC	LWCCN [%]	LWCCL [%]	SW	SW-rnd	CC	CC-rnd	*A*
OMN	1	100	100	2.55	2.22	0.15	0.028	0.023
OCN	1	100	100	9.51	9.74	0.007	0.00021	-0.158
OSN	36	93.35	94.11	11.8	11.74	0.001	0.00017	-0.171

Cyclic dependencies

- **A giant SCC in the SWEET ontology OMN**
  - large cyclic dependencies among SWEET modules
  - small link reciprocity, but a large path reciprocity
    → cyclic dependencies among SWEET modules are mostly indirect

- A large number of small-size SCCS in the SWEET OCN, large cyclic dependencies among SWEET classes are absent

- Two classes involved in mutual subsumption relations

Table 4.4: Strongly connected components in the SWEET ontology networks. #SCC – the total number of strongly connected components, LSCCN – the percentage of nodes in the largest SCC, LSCCL – the percentage of links in the largest SCC, S – the percentage of nodes contained in all SCCs, R – link reciprocity, R_n – normalized link reciprocity, R_p – path reciprocity, C – the percentage of SCCs that are pure cycles.

Network	#SCC	LSCCN [%]	LSCCL [%]	S [%]	R	R_n	R_p	C [%]
OMN	3	61.57	60.63	64.53	0.0545	0.0275	0.608	33.33
OCN	410	0.17	0.20	15.05	0.1214	0.1212	0.0136	80.24
OSN	1	0.03	0.03	0.03	0.0004	0.0003	0.0001	100

- **Metric-based comparison test to determine the differences between modules in the giant SCC and the rest of SWEET modules**
- SWEET has a strongly connected core encompassing the most reused and the most important SWEET modules

Metric	Avg(GSCC)	Avg(Rest)	U	p	NullHyp	PS_1	PS_2
LOC	106.8	101.1	6029	0.0045	rejected	0.61	0.38
TEXPR	5.26	4.11	5412	0.1872	**accepted**	0.5	0.39
AEXPR	0.068	0.071	5058	0.6522	**accepted**	0.49	0.45
AXM	92.8	87.3	6033	0.0044	rejected	0.61	0.38
HVOL	2905	2855.5	6048	0.0039	rejected	0.62	0.38
HDIF	20.5	17.7	6376	0.0002	rejected	0.65	0.35
NCLASS	34.06	27.04	5773	0.0274	rejected	0.58	0.4
NINST	9.24	13.97	5007	0.7448	**accepted**	0.24	0.27
IN	8.34	1.22	9110	$< 10^{-4}$	rejected	**0.91**	0.04
OUT	5.77	5.35	5002	0.7541	**accepted**	0.47	0.46
TOT	14.1	6.55	8057	$< 10^{-4}$	rejected	**0.81**	0.15
BET	870.8	20.6	8781	$< 10^{-4}$	rejected	**0.89**	0.09
PR	0.0066	0.0022	8971	$< 10^{-4}$	rejected	**0.92**	0.08
HITSH	0.0642	0.0467	6048	0.0039	rejected	0.62	0.38
HITSA	0.0549	0.0064	9414	$< 10^{-4}$	rejected	**0.97**	0.03
HK	717545.28	7888.64	8959	$< 10^{-4}$	rejected	**0.92**	0.08
AP	1.74	1.28	4892	0.9666	**accepted**	0.24	0.23
CR	0.11	0.09	5104	0.5729	**accepted**	0.3	0.25
RR	0.23	0.23	5025	0.7125	**accepted**	0.5	0.47
NEC	5.12	4.68	4962	0.8298	**accepted**	0.46	0.44
REC	9.49	8.76	4981	0.7946	**accepted**	0.47	0.49
CON	0.21	0.22	5470	0.1438	**accepted**	0.44	0.56
EXP	0.29	0.31	5498	0.1259	**accepted**	0.43	0.56
CUTR	0.000027	0.000029	5497	0.1266	**accepted**	0.44	0.56
AVGODF	0.24	0.27	5471	0.1432	**accepted**	0.44	0.56
MAXODF	0.94	0.95	4946	0.8615	**accepted**	0.08	0.1
FODF	0.76	0.73	5395	0.2015	**accepted**	0.55	0.44

- **Degree distribution analysis:** the SWEET ontology networks contain hubs (highly coupled nodes)
- **Metric-based comparison test:** hubs tend to more voluminous and more functionally important modules than non-hub modules

Metric	Avg(Hubs)	Avg(Rest)	U	p	NullHyp	PS_1	PS_2
LOC	138.4	93	6185	$< 10^{-4}$	rejected	**0.79**	0.21
TEXPR	7.6	3.9	5449	$< 10^{-4}$	rejected	0.66	0.27
AEXPR	0.076	0.068	4579	0.07	**accepted**	0.57	0.4
AXM	122.5	79.7	6097	$< 10^{-4}$	rejected	**0.77**	0.22
HVOL	3931.3	2526.1	6237	$< 10^{-4}$	rejected	**0.79**	0.21
HDIF	23.1	18.2	5797	$< 10^{-4}$	rejected	0.74	0.26
NCLASS	46.6	26.1	5947	$< 10^{-4}$	rejected	**0.75**	0.24
NINST	8.8	11.8	4316	0.28	**accepted**	0.19	0.29
IN	14.7	2.5	7214	$< 10^{-4}$	rejected	**0.9**	0.06
OUT	7.4	5	5175	0.0006	rejected	0.62	0.31
BET	1438.7	236.01	6815	$< 10^{-4}$	rejected	**0.87**	0.13
PR	0.0128	0.0022	6737	$< 10^{-4}$	rejected	**0.86**	0.14
HITSA	0.09	0.01	7326	$< 10^{-4}$	rejected	**0.93**	0.07
HITSH	0.08	0.04	5435	$< 10^{-4}$	rejected	0.69	0.31
HK	1681429.9	19033.9	7688	$< 10^{-4}$	rejected	**0.98**	0.02
AP	0.83	1.82	4198	0.45	**accepted**	0.19	0.26
CR	0.07	0.11	3926	0.99	**accepted**	0.29	0.28
RR	0.28	0.22	4845	0.01	rejected	0.61	0.38
NEC	7	4.2	5257	0.0003	rejected	0.63	0.29
REC	13.2	7.8	5019	0.003	rejected	0.62	0.34
CON	0.19	0.22	4450	0.15	**accepted**	0.44	0.57
EXP	0.26	0.31	4409	0.18	**accepted**	0.44	0.56
CUTR	0.000024	0.000029	4404	0.19	**accepted**	0.44	0.56
AVGODF	0.22	0.26	4359	0.24	**accepted**	0.44	0.55
MAXODF	0.96	0.94	4042	0.75	**accepted**	0.09	0.07
FODF	0.78	0.74	4245	0.38	**accepted**	0.53	0.45

GCE metrics as ontology metrics

- M - an ontology module within a modularized ontology
- G(M) - a graph showing dependencies among classes in M
  - G(M) is a subgraph of the ontology class network

- Basic ontology module cohesion metrics ignoring external dependencies
  - DEN - the density of G(M)
  - COMP - the number of weakly connected components in G(M)

Table 4.14: The values of the Spearman correlation coefficient for GCE metrics and metrics of internal ontology module density (DEN) and connectedness (COMP).

	EXP	CON	CUTR	AODF	FODF
DEN	-0.035	-0.056	-0.038	-0.109	0.076
COMP	0.174	0.265	0.175	0.253	-0.235

- **Weak correlations → ontology cohesion metrics based solely on internal class dependencies are unable to identify modules whose constituent classes form strong clusters in the OCN**

Cohesion of SWEET modules

- **SWEET ontology modules has a satisfactory degree of cohesion**
  - 18 modules (8.87%) are Radicchi strong clusters
  - 195 modules (96.08%) are Radicchi weak clusters
  - only 8 modules are poorly cohesive (non-Radicchi-weak clusters)
  - **poorly cohesive modules have a low centrality in the OMN**

Module	LOC	TEXPR	IN	OUT	PR	BET	CON
stateSpaceConfiguration.owl	106	0	1	2	0.0016	12	0.75
stateTimeFrequency.owl	72	0	2	7	0.0021	260	0.75
quanTimeAverage.owl	89	1	3	8	0.0012	451	0.74
stateSpace.owl	70	0	0	5	0.0010	0	0.65
realmAtmoWeather.owl	61	4	0	7	0.0010	0	0.65
reprSpaceDirection.owl	97	0	9	2	0.0045	16	0.61
phenOcean.owl	15	1	2	2	0.0014	13	0.6
stateTime.owl	83	5	3	5	0.0015	7	0.5
A	104.6	4.82	5.6	5.6	0.0049	544	0.22

Outline

- Introduction
- Fundamentals of complex network analysis
- Methods for annotated networks
- Case study 1 — analysis of enriched co-authorship networks
- Case study 2 — analysis of enriched ontology networks
- **Conclusions**

Conclusions

- Methods to analyze annotated social and information networks focused on categorically induced subgraphs, block models and attachment preferences

- Case studies related to analysis annotated networks with numeric attributed being domain-dependent metrics

  - **analysis of enriched co-authorship networks**
    - an in-depth evaluation of research collaboration and mutual relationships between collaboration and other determinants of research performance

  - **analysis of enriched ontology networks**
    - evaluation of design quality of modular ontologies with respect to modular design principles originating from software engineering

- More about the topics of the tutorial including presented case studies can be found in