InfoLab

IC612: Data Warehousing and Data Mining

Lecture 12:
Social Network Analysis

Min-Soo Kim

DGIST 대구경북과학기술원
Daegu Gyeongbuk Institute of Science & Technology

Concepts and Terms

- **Homophily: nodes that are connected to one another are more likely to have similar properties**
  - e.g., a person's friendship links in Facebook may be drawn from previous acquaintances in school and work
  - e.g., the friendship links may often imply common interests between the two parties
  - c.f., *Birds of a feather flock together*

- **Triadic closure (triangle)**
  - if two individuals in a social network have a friend in common, then it is more likely that they are either connected or will eventually become connected in the future
  - structural version of homophily
    - c.f., homphily is typically exhibited in terms of node attributes

Clustering coefficient

- **Measure of the inherent tendency of a network to cluster**
- **$S_i \subseteq N$: the set of nodes connected to node $i \in N$**
  - in the undirected network G = (N,A)
- **n_i : the cardinality of S_i**

$$\eta(i) = \frac{|\{(j,k) \in A : j \in S_i, k \in S_i\}|}{\binom{n_i}{2}}$$

- **Average clustering coefficient : average value of η(i) over all nodes in the network**
  - triadic closure property increases the clustering coefficient of real-world networks

Dynamics of network formation

- **Networks are continuously growing over time with new nodes and edges being added constantly**
  - e.g., World Wide Web, social networks

1. **Preferential attachment**
2. **Small world property**
3. **Densification**
4. **Shrinking diameters**
5. **Giant connected component**

Preferential attachment

- **Likelihood of a node receiving new edges increases with its degree**
  - highly connected individuals will typically find it easier to make new connections
- **π(i) : probability that a newly added node attaches itself to an existing node i in the network**

$$\pi(i) \propto \mathrm{Degree}(i)^{\alpha}$$

  - parameter α : dependent on the domain of the network

- **Scale-free network (Barabasi-Albert model)**
  - $\alpha \approx 1$: the proportionality is linear
  - e.g., World Wide Web, social networks, and biological networks

Small world property

- **Most real networks are assumed to be “small world”**
  - average path length between any pair of nodes is quite small
- **Milgram’s experiment in the sixties conjectured**
  - the distance between any pair of nodes is about six
- **Network containing n(t) nodes at time t**
  - many models postulate that the average path lengths grow as log(n(t))
  - a small number, even for very large networks

Densification

- **Almost all real-world networks add more nodes and edges over time than are deleted**
- **The impact of adding new edges generally dominates the impact of adding new nodes**
  - this implies that the graphs gradually densify over time
  - # edges growing superlinearly with # nodes
- **Densification power law**

$$e(t) \propto n(t)^{\beta}$$

  - exponent β : a value between 1 and 2
  - β = 1 : the average degree of the nodes is not affected by the growth of the network
  - β = 2 : the total number of edges e(t) remains a constant fraction of the complete graph of n(t) nodes as n(t) increases

Shrinking diameters

- **As the network densifies, the average distances between the nodes shrink over time**

- **Contrast to conventional models that suggest that the diameters should increase as log(n(t))**
  - consequence of the fact that the addition of new edges dominates the addition of new nodes
  - if the impact of adding new nodes were to dominate, then the average distances between nodes would increase over time

Giant connected component

- **As the network densifies over time, a giant connected component emerges**

- **It is consistent with the principle of preferential attachment**
  - newly incoming edges are more likely to attach themselves to the densely connected and high-degree nodes in the network

- **It has a confounding impact on network clustering algorithms**
  - it typically leads to unbalanced clusters, unless the algorithms are carefully designed

Power-Law degree distributions

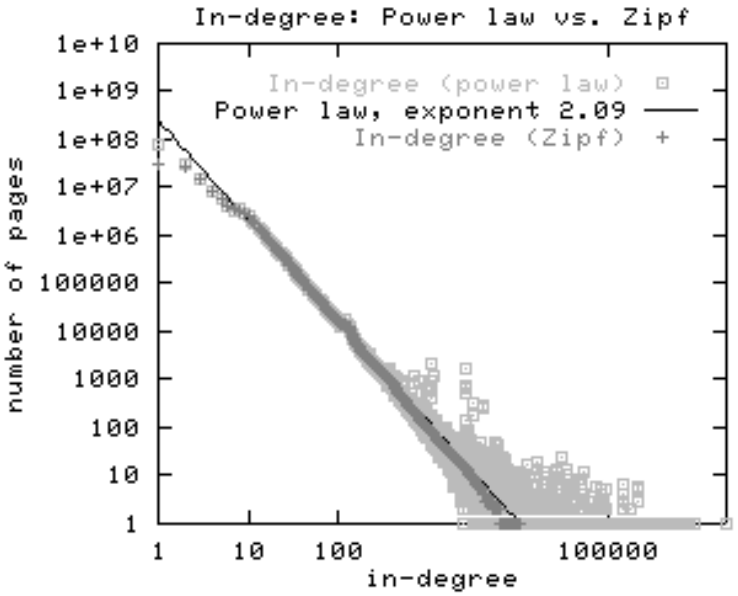
- **Consequence of preferential attachment**
  - a small minority of high-degree nodes continue to attract most of the newly added nodes

- **Power-law degree distribution**
  - number of nodes P(k) with degree k is regulated by

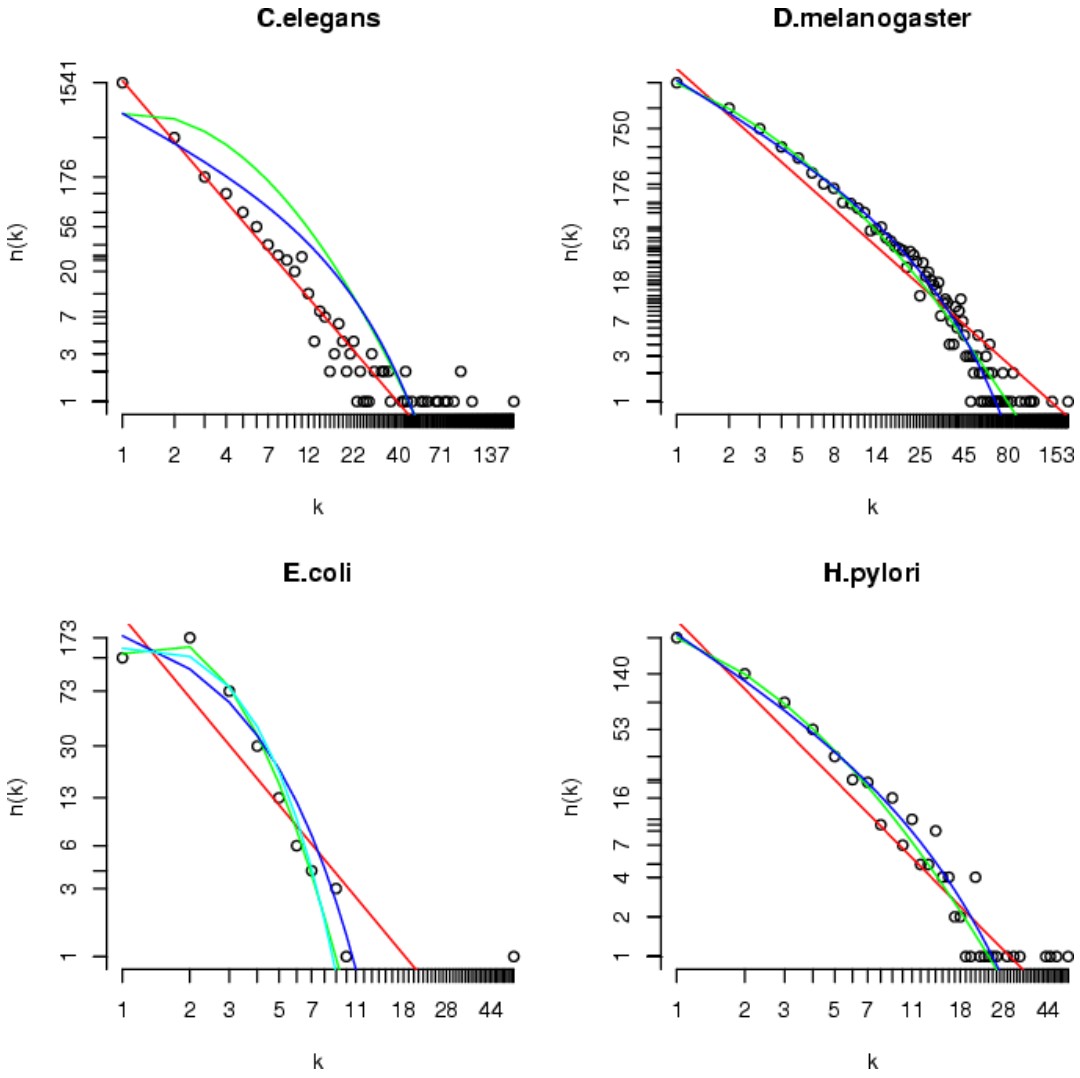
$$P(k) \propto k^{-\gamma}$$

  - parameter γ : between 2 and 3
  - the larger values of γ lead to more small degree nodes
  - e.g., $\gamma = 3$, the vast majority of the nodes in the network will have a degree of 1

Examples of power-law degree dist.



web structure



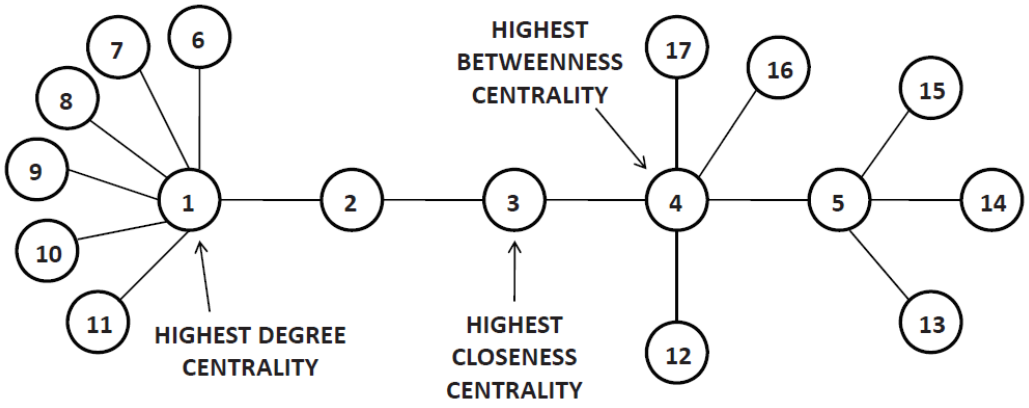
red line is power-law

Degree Centrality

- **Degree centrality $C_{D(i)}$ of a node i (undirected network)**

$$C_D(i) = \frac{\text{Degree}(i)}{n-1}$$

  - hub nodes tend to be more central to the network
  - drawback: consider nodes beyond the immediate neighborhood
  - e.g., node 1 in is highest degree, but cannot be viewed as central

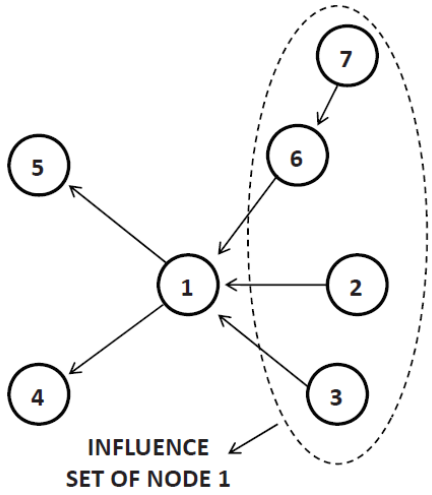


InfoLab DGIST 대구경북과학기술원

Prestige

- **Degree prestige (directed networks only)**
  - idea : only indegree contributes to prestige (similar to PageRank)
  - e.g., node 1

$$P_D(i) = \frac{\text{Indegree}(i)}{n-1}$$



InfoLab DGIST 대구경북과학기술원

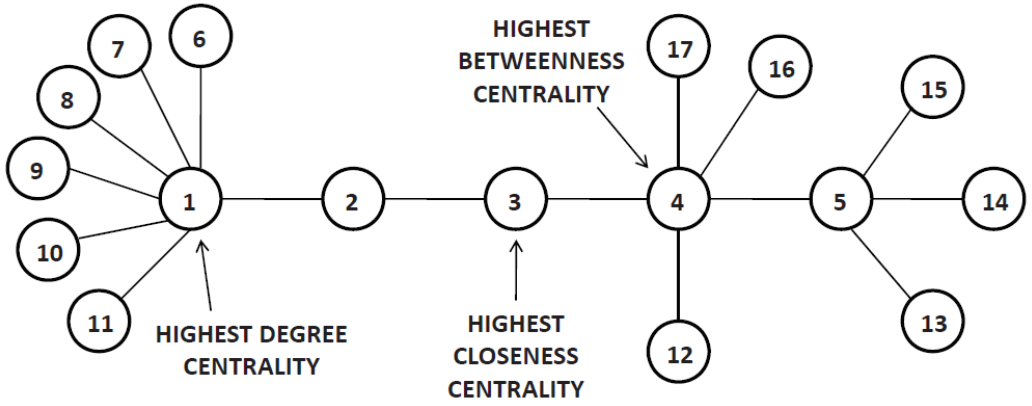
Closeness Centrality

- **Notion of closeness centrality is meaningfully defined w.r.t undirected and connected networks**
  - AvDist(i): average shortest path distance, starting from node i
  - Dist(i, j): the shortest path distance between node i and node j

$$\mathrm{AvDist}(i) = \frac{\sum_{j=1}^{n} \mathrm{Dist}(i,j)}{n-1}$$

  - closeness centrality : inverse of AvDist(i)
    - between 0 and 1 since $1 \leq \mathrm{AvDist}(i)$

$$C_C(i) = 1/\mathrm{AvDist}(i)$$



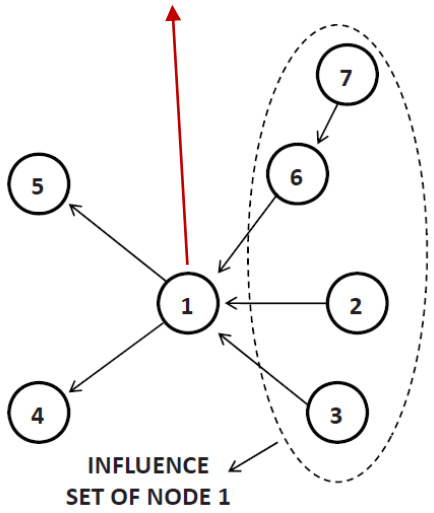
InfoLab DGIST 대구경북과학기술원

Proximity prestige

- **Used to measure prestige in directed networks**
  - confounding factor: directed paths may not exist from other nodes to node i (e.g., node 7)
- **Influence(i) : the set of nodes that can reach node i with a directed path**
  - e.g., all recursively defined followers of node i in Twitter

proximity prestige of node 1

$$\frac{4}{6} \times \frac{4}{5} = \frac{16}{30}$$



$$\text{InfluenceFraction}(i) = \frac{|\text{Influence}(i)|}{n-1}$$

$$\text{AvDist}(i) = \frac{\sum_{j \in \text{Influence}(i)} \text{Dist}(j,i)}{|\text{Influence}(i)|}$$

$$P_P(i) = \frac{\text{InfluenceFraction}(i)}{\text{AvDist}(i)}$$

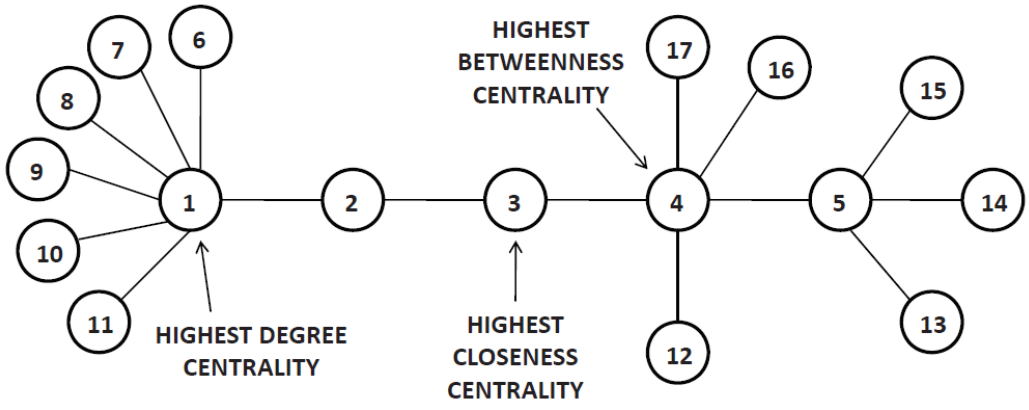
(between 0 and 1)

InfoLab DGIST 대구경북과학기술원

Betweenness centrality

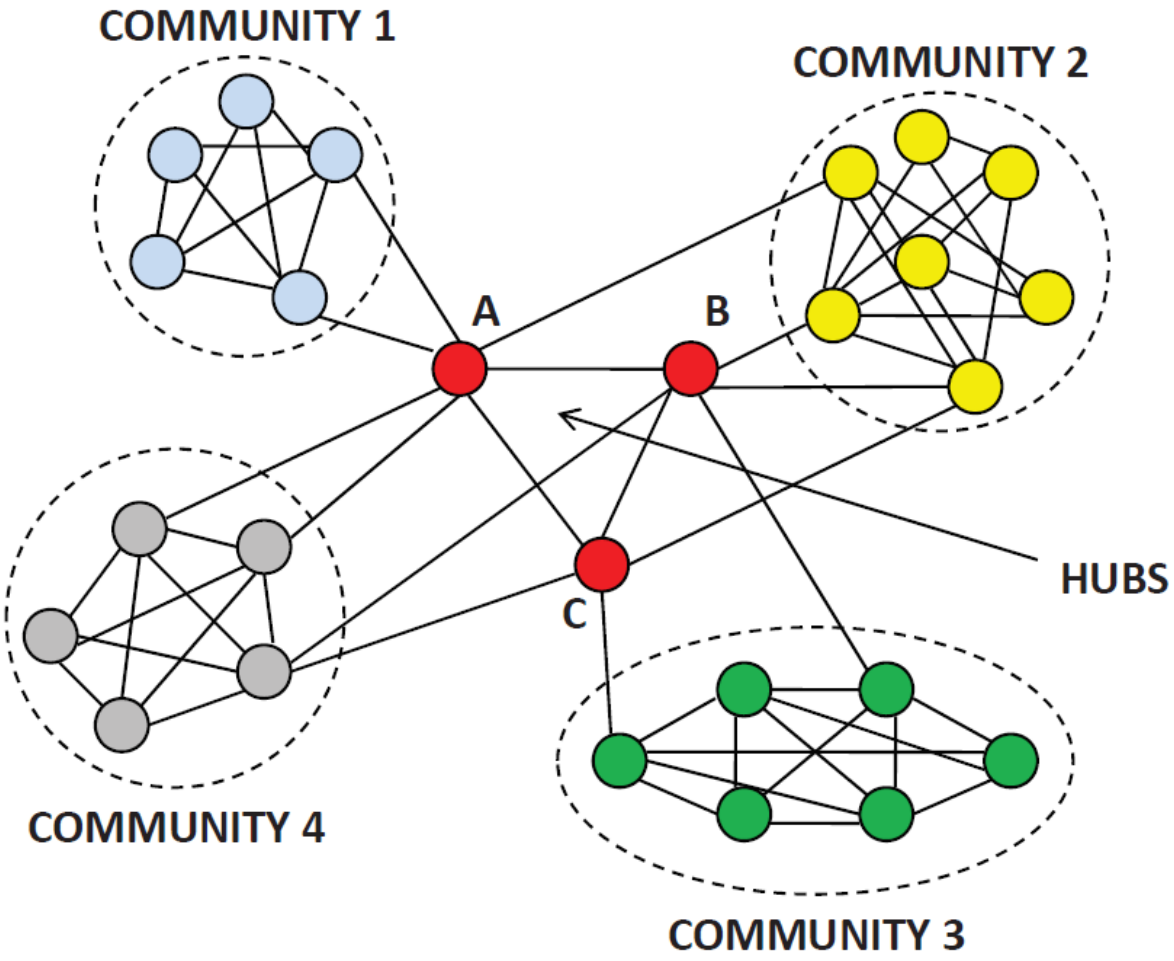
- **Closeness centrality does not account for the criticality of the node**
  - especially, in terms of # shortest paths passing through it
  - e.g., node 3 (highest closeness centrality) is not highest critical
    - node 4 controls the flow of information between nodes 12 and 17
    - node 3 does not control it
- **$f_{jk}(i) = q_{jk}(i) / q_{jk}$**
  - q_{jk}: # shortest paths between nodes j and k
  - $q_{jk}(i)$: # shortest paths between nodes j and k through node i

$$C_B(i) = \frac{\sum_{j<k} f_{jk}(i)}{\binom{n}{2}}$$



InfoLab DGIST 대구경북과학기술원

Betweenness centrality for edges



Community detection

- **"community detection" is an approximate synonym for "clustering" in the context of social network analysis**
- **Clustering of networks and graphs is also sometimes referred to as "graph partitioning"**
- **Difficulty in cleanly separating out different clusters for social networks**
  - different parts of the social network have different edge densities
  - real social networks often have a giant component that is densely connected
    - partition balancing : different clusters have similar numbers of nodes

Girvan-Newman Algorithm

- **Concept: edges with high betweenness have a tendency to connect different clusters**
  - disconnection of these edges will result in a set of connected components that corresponds to the natural clusters
  - top-down hierarchical clustering algorithm that creates clusters by successively removing edges with the highest betweenness
    - until the graph is disconnected into required # connected components

- **Main challenge: computing edge betweenness values**

- **Using edge lengths c_{ij}, rather than edge weights w_{ij}**
  - edge lengths may be viewed as in the inverse of the edge weights

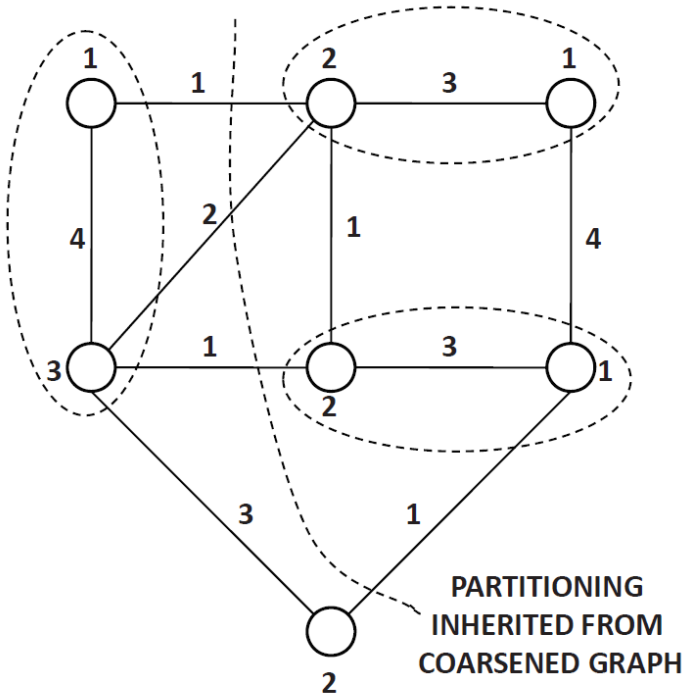
Algorithm *GirvanNewman*(Graph: $G = (N, A)$, Number of Clusters: k, Edge lengths: $[c_{ij}]$)
begin
  Compute betweenness value of all edges in graph G;
  repeat
    Remove edge (i, j) from G with highest betweenness;
    Recompute betweenness of edges affected by removal of (i, j);
  until G has k components remaining;
  return connected components of G;
end

Multilevel Graph Partitioning: METIS

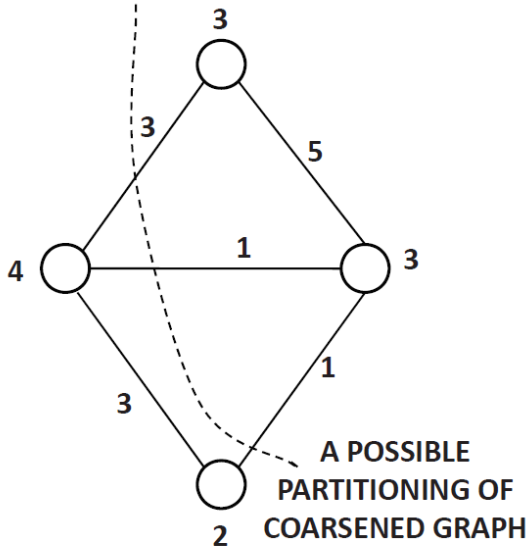
- **Can be used to perform either k-way partitioning or 2-way partitioning**

- **Principle: partitioning of a coarsened representation of a graph (approximate partition of the original graph)**
  - coarsened representation: obtained by contracting some of the adjacent nodes into a single node
  - contraction may result in collapsed edges
  - weights of the contracted nodes are equal to the sum of the weights of the constituent nodes in the original graph
  - the edges across contracted nodes are consolidated into a single edge with the weights of the constituent edges added together
  - resulting partition can be refined with an algorithm, such as the Kernighan-Lin algorithm

- **Allow the specification of weights on both the nodes and edges in the clustering process**
  - w_{ij}: edge weight
  - v_i: node weight



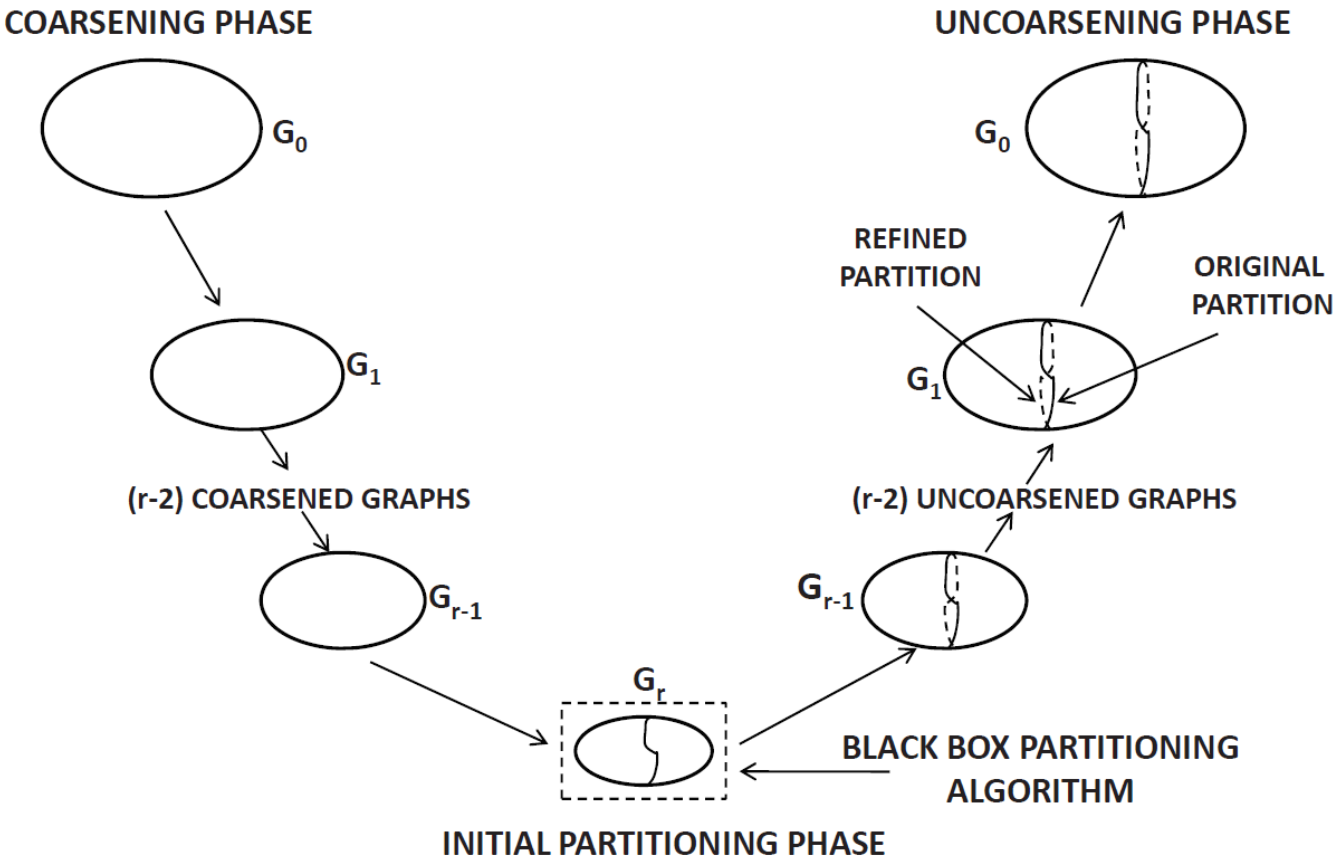
(a) Original graph with inherited partition from coarsened graph



(b) Coarsened graph with partition

InfoLab DGIST 대구경북과학기술원

- **Multilevel scheme enhances this basic approach with multiple levels of coarsening and refinement**
  - ➢ to obtain a good trade-off between quality and efficiency



InfoLab DGIST 대구경북과학기술원

Multilevel partitioning scheme uses three phases

1. Coarsening phase

- for original graph $G = G_0$, create a sequence of successively smaller graphs, $G_0, G_1, G_2 \ldots G_r$
- when coarsening G_{m-1} to G_m, small sets of non-overlapping and tightly interconnected nodes are identified and contracted
- final graph G_r is typically smaller than a hundred nodes

2. Partitioning phase

- any off-the-shelf algorithm can be used to create a high-quality balanced partitioning from graph Gr

3. Uncoarsening phase (refinement)

- the graphs are expanded back to their successively larger versions $G_r, G_{r-1} \ldots G_0$
- graph G_{m-2} inherits the refined partition from G_{m-1}

일정	12월 02일(금)			
09:30~11:30(120')	**포스터세션** [E7 로비]			
11:30~12:30(60')	**기조강연 II** : Prof. Bin Yu [E1 컨벤션홀 / 좌장 : 문태섭 교수]			
12:30~14:00(90')	**오찬** [E1 정찬실]			
세션	**에너지시스템공학** [E7 241]	**정보통신융합공학** [E7 236]	**뉴바이올로지** [E7 233]	**융복합대학** [E7 L29]
14:00~14:40(40')	**Nenad Markovic** (Argonne National Laboratory) * 좌장 : 유종성(DGIST)	**Xiaogang Wang** (Chinese University of Hong Kong) * 좌장 : 김민수(DGIST)	**Richard Zare** (Stanford Univ.) * 좌장 : 남홍길(DGIST)	**환영사** (14:00~14:20)
14:40~15:20(40')	**이종호** (한국과학기술연구원) * 좌장 : 유종성(DGIST)	**최승진** (POSTECH) * 좌장 : 김민수(DGIST)	**이종봉** (POSTECH) * 좌장 : 남홍길(DGIST)	**융복합대학 소개** (14:20~15:00) 휴식(15:00~15:20)
15:20~15:50(30')	휴식			**UGRP 소개** (15:20~15:50)
15:50~16:30(40')	**Samuel Mao** (UC Berkeley) * 좌장 : 김하석(DGIST)	**Ichiro Takeuchi** (Nagoya Institute of Technology) * 좌장 : 김민수(DGIST)	**Nobuyoshi Hirose** (Keio Univ.) * 좌장 : 남홍길(DGIST)	**FGLP 소개 및 Q&A** (15:50~17:30)
16:30~17:10(40')	**Piotr Zelenay** (Los Alamos National Laboratory) * 좌장 : 김하석(DGIST)	**이경무** (서울대) * 좌장 : 김민수(DGIST)	**이서구** (연세대 의생명연구원) * 좌장 : 남홍길(DGIST)	
17:10~17:50(40')	**유종성** (DGIST) * 좌장 : 김하석(DGIST)	**문태섭** (DGIST) * 좌장 : 김민수(DGIST)	**박상철** (DGIST) * 좌장 : 남홍길(DGIST)	-
17:50~18:20(30')	휴식 및 이동			
18:20~20:00(100')	**만찬** [E8 학술정보관 로비]			

Understanding Deep Learning and Neural Semantics

- **Xiaogang Wang (Chinese University of Hong Kong)**

Deep learning has achieved great success in computer vision. Many people believe that the success is due to employing a huge number of parameters to fit big training data. In this talk, I will show that neuron responses of deep models have clear semantic interpretation, which is supported by our research on multiple fields of face recognition, object tracking, human pose estimation, and crowd video analysis. In particular, the responses of neurons in the top layers have sparseness and strong selectiveness object classes, attributes and identities. Sparseness and selectiveness are strongly correlated. Such selectiveness is naturally obtained through large scale training without adding extra regularization during the training process. By understanding neural semantics, we are inspired to develop new network architectures and training strategies and they effectively improve a broad range of applications in face recognition, face detection, compressing neural networks, object tracking, learned structured feature representation in human pose estimation, and effectively learning dynamic feature representations of different semantic units in video understanding.

Bayesian Deep Probabilistic Models

- 최승진 (POSTECH)

Deep generative models are a rich class of models for density estimation which specify a generative process for observed data using a set of stochastic latent variables. Among those, of particular interest is the variational autoencoder (VAE) which is a deep directed generative model with continuous latent variables where a pair of probabilistic encoder (bottom-up) and decoder (top-down) is jointly learned by stochastic gradient variational Bayes (SGVB). In this talk, I begin with introducing our recent elaboration of Gaussian VAE, approximating the local covariance matrix of the decoder as an outer product of the principal direction at a position determined by a sample drawn from Gaussian distribution. I show that this model, referred to as VAE-ROC, better captures the data manifold, compared to the standard Gaussian VAE where independent multivariate Gaussian was used to model the decoder. Then, I introduce an extension of VAE-ROC to handle mixed ordinal and continuous data, where Gaussian copula is adopted to model the local dependency in mixed ordinal and continuous data, leading to "Gaussian copula variational autoencoder"

Quickly Analyzing Sensitivity of Incremental Data Update for Big Data Machine Learning in Changing Environment

- **Ichiro Takeuchi (Nagoya Institute of Technology)**

In this talk, we are concerned with large-scale machine learning problems in changing environment where a small part of the dataset which have been used for training a machine learning model is incrementally updated, and the effect of the data update must be quickly incorporated into the machine learning model. When the entire dataset is large, even if the amount of the data update is fairly small, the computational cost of re-training the machine learning model would be prohibitively large. In this talk, we introduce a novel method called Quick Sensitivity Analysis (QSA) for efficiently incorporating such a data update effect into the machine learning model without spending large computational cost for re-training it. The proposed QSA method provides bounds on the unknown yet-to-be trained machine learning model with the cost only proportional to the size of the data update. We demonstrate through numerical experiments that the proposed QSA method provides sufficiently tight bounds with negligible computational costs especially when a small part of the dataset is updated in a large-scale machine learning problem.

Deep Image Restoration

- 이경무 (서울대)

Image restoration or deconvolution that is to recover the original clean image from a noisy or corrupted image has long been an important and fundamental problem in image processing and computer vision. In this talk, a very effective deep learning-based image restoration techniques will be addressed, which are based on very deep convolutional networks. Increasing the network depth enlarges the receptive field, resulting a significant improvement in accuracy. By cascading small filters many times in deep network structures, contextual information over large image regions is exploited in an efficient way. With very deep networks, however, convergence speed becomes a critical issue during training. Simple yet effective training procedures for successful convergence of very deep networks will be discussed including residual learning and adjustable gradient clipping. We show that our proposed methods outperform existing deconvolution methods quantitatively as well as qualitatively.

Neural Universal Discrete Denoiser

- 문태섭 (DGIST)

In this talk, I will present a novel framework of applying deep neural network (DNN) to discrete denoising problem. DNN has recently shown remarkable performance improvements in diverse applications, and most of the success are based on the supervised learning framework. While successful in many applications, it is not straightforward to apply such framework to the discrete denoising problem, in which a denoiser tries to estimate an unknown finite-valued clean data based on its noisy observation. The reason is because the ground-truth label for a denoiser is the clean data subject to the estimation and is clearly not available for training a denoiser. In this talk, I follow the framework of DUDE (Discrete Universal DEnoiser) and devise a novel way of training a DNN as a discrete denoiser without any ground-truth labels. The key idea is to define "pseudo-labels" based on an unbiased estimate of the true loss of a denoiser and use them as targets for DNN parameters to optimize for. The resulting scheme is dubbed as Neural DUDE, and our experiments show that Neural DUDE significantly outperforms the original DUDE, which is the state-of-the-art on several discrete denoising problems. Furthermore, I will show that Neural DUDE overcomes the critical limitation of DUDE, namely, it is much more robust to the choice of the hyper-parameter and has a concrete way of choosing the best hyper-parameter for given data. Such property becomes an attractive feature of Neural DUDE in practice. Finally, I will conclude with some potential future research directions, such as extending the framework to the denoising of continuous-valued data.

InfoLab

Thank you!

DGIST 대구경북과학기술원
Daegu Gyeongbuk Institute of Science & Technology