

Network structure identification

Network Application Diagnostics
B2M32DSA

Radek Mařík

Czech Technical University
Faculty of Electrical Engineering
Department of Telecommunication Engineering
Prague CZ

October 16, 2025



1 Node Influence

- Node Roles
- Hubs and Authorities

2 Network Clustering

- Multidimensional Clustering
- Clustering Coefficients
- Graph Laplacian



Outline

1 Node Influence

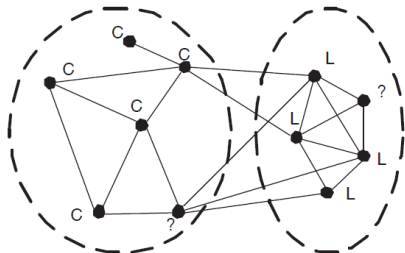
- Node Roles
- Hubs and Authorities

2 Network Clustering

- Multidimensional Clustering
- Clustering Coefficients
- Graph Laplacian



Node Roles and Community Structure ^[STE07]



predicting the political leaning of a person, either conservative (C) or liberal (L)

- *influence maximization* can be thought of as finding the best k people to target in order to maximize the number of people that will eventually be influenced
- *link-based classification* is the task of categorizing nodes using the node features and its link information

Community Metric - Prerequisites ^[STE07]

- *Assumption*: a community is defined by a clique (maximal complete subgraph) in a network.
- **rawComm** is to be an approximate measure of the number of communities to which a node is attached.
- *incomplete edge* ... an edge that connects two nodes in different communities
- *impure edge* ... a non-link that appears within community.
- p ... the probability that two linked nodes are in the same community

$$p = \frac{\text{Complete node pairs}}{\text{Total linked node pairs}}$$

- q ... the probability that two non-linked nodes are in different communities.

$$q = \frac{\text{Pure node pairs}}{\text{Total non-linked node pairs}}$$



Community Metric ^[STE07]

- **rawComm** metric

$$\text{rawComm} = \sum_{v \in N(u)} \tau_u(v)$$

where

- $N(u)$... the neighborhood of u ,
that is all of the nodes that are directly linked to u

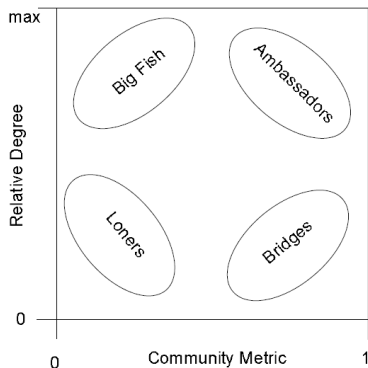
-

$$\tau_u(v_i) = \frac{1}{1 + \sum_{v_j \in N(u)} I(v_i, v_j) \cdot p + \bar{I}(v_i, v_j) \cdot (1 - q)}$$

- $I(x, y)$... an indicator function that is 1 if there is a link between x and y and 0 otherwise. $\bar{I}$ is 1 if there is not a link and 0 otherwise.
- The denominator in the definition of τ is the expected number of other nodes in u 's neighborhood are in a community with v_i .
 - The 1 represents the node v_i itself.
 - $I(v_i, v_j) \cdot p + \bar{I}(v_i, v_j) \cdot (1 - q)$... the probability of v_i and v_j being in the same community.



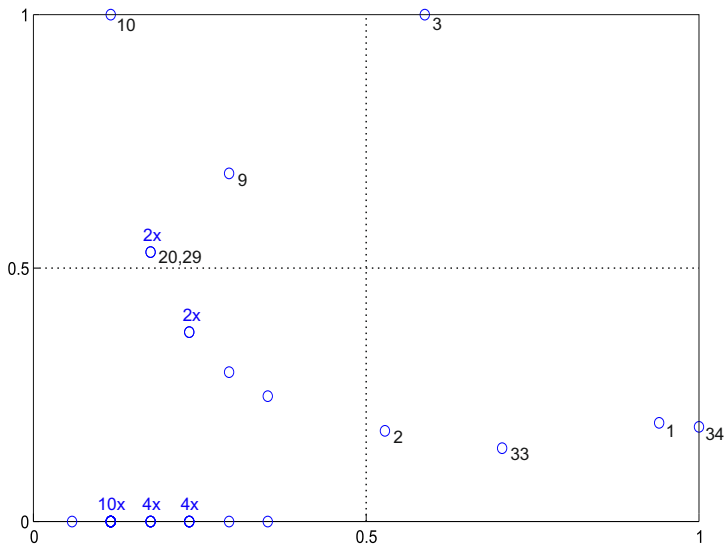
Community-based Node Roles ^[STE07]



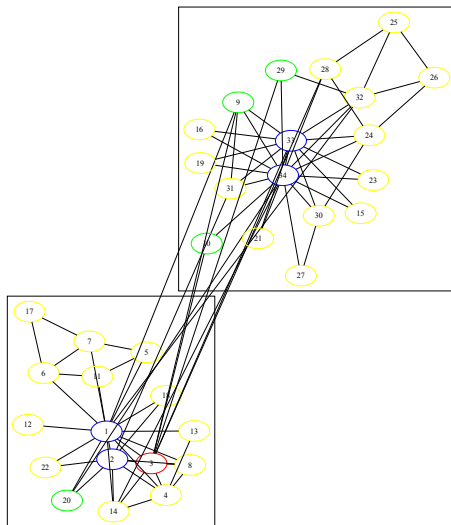
- a *authority* how much knowledge, information, etc. held by a node on a topic.
- a *hub* how well a node 'knows' where to find information on a given topic.
- an *ambassador* has links to many nodes from different communities
- a *big fish* has links only to other nodes in the same community
- a *bridge* because they serve as bridges between a small number of communities
- the *loners* ... those with a low relative degree and low community



Community-based Node Roles - Karate Club



Community-based Node Roles - Karate Club



Outline

1 Node Influence

- Node Roles
- Hubs and Authorities

2 Network Clustering

- Multidimensional Clustering
- Clustering Coefficients
- Graph Laplacian



Hubs and Authorities I [Kle98, Kle99, New10]

- Proposed by Kleinberg in 1998 ^[Kle98]
- In some networks it is appropriate also to accord a vertex high centrality if it points to others with high centrality
 - papers and their reviews
 - it makes sense only in directed networks (direction)
- **Authorities** are nodes that contain useful information on a topic of interest.
- **Hubs** are nodes that tell us where the best authorities are to be found.
- The centrality algorithm is called **hyperlink-induced topic search** or **HITS**
- Each vertex i has an authority centrality x_i and a hub centrality y_i



Hubs and Authorities II [Kle98, Kle99, New10]

- The authority centrality of a vertex is defined to be proportional to the sum of the hub centralities of the vertices that point to it:

$$x_i = \alpha \sum_j A_{ij} y_j$$

where α is a constant.

- The hub centrality of a vertex is proportional to the sum of the authority centralities of the vertices it points to:

$$y_i = \beta \sum_j A_{ji} x_j$$

where β is a constant.

- In matrix terms

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{y} \qquad \mathbf{y} = \beta \mathbf{A} \mathbf{x}$$

- Combining both and setting $\gamma = (\alpha\beta)^{-1}$

$$\mathbf{A} \mathbf{A}^T \mathbf{x} = \gamma \mathbf{x} \qquad \mathbf{A}^T \mathbf{A} \mathbf{y} = \gamma \mathbf{y}$$



Hubs and Authorities III

[Kle98, Kle99, New10]

$$\mathbf{A}\mathbf{A}^T\mathbf{x} = \gamma\mathbf{x}$$

$$\mathbf{A}^T\mathbf{A}\mathbf{y} = \gamma\mathbf{y}$$

- the authority and hub centralities are respectively given by eigenvectors of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ with the same eigenvalue γ .
 - The same eigenvalue γ ?

$$\mathbf{A}\mathbf{A}^T\mathbf{x} = \gamma\mathbf{x} \quad \dots | \mathbf{A}^T \times \quad (1)$$

$$\mathbf{A}^T\mathbf{A}(\mathbf{A}^T\mathbf{x}) = \gamma(\mathbf{A}^T\mathbf{x}) \quad (2)$$

$$(3)$$

- The relation between both centralities

$$\mathbf{y} = \mathbf{A}^T\mathbf{x}$$

- $\mathbf{A}\mathbf{A}^T$ is the cocitation matrix.
- $\mathbf{A}^T\mathbf{A}$ is the bibliographic coupling matrix.
- HITS does not struggle with zero centralities outside strong components.



Outline

1 Node Influence

- Node Roles
- Hubs and Authorities

2 Network Clustering

- Multidimensional Clustering
- Clustering Coefficients
- Graph Laplacian



Data Clustering [XW05, EK10]

- Data are **classified** or grouped into a set of categories or clusters.
 - **supervised** ... with a teacher
 - a finite set of class/category labels/tags is provided
 - **unsupervised** ... without a teacher
 - based on similarities of objects
- A **cluster** is a collection of objects that are similar to each other using some attribute.
- A cluster of objects can be treated as a group.
- Let $P = \{p_1, \dots, p_N\}$ be a set of N data points representing N objects.
- The goal of **clustering** (CZ shlukování) is to divide P into K groups $C_1, \dots, C_K$ so that data belonging to a group are more similar to each other than data from different groups.
- Each C_i is called a **cluster** (CZ shluk).
- Each object p_j is described by a vector $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jd})^T \in \mathbb{R}^d$ and each measure x_{ji} is called to be a **feature** (attribute, dimension, or variable) (CZ příznak).



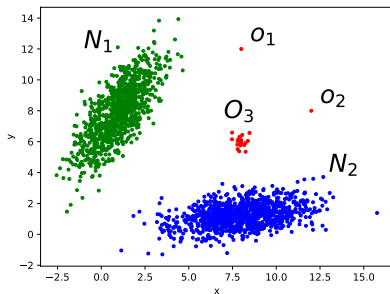
Input Data ^[Agg17]

- Record data
 - Univariate
 - Multivariate
- Attributes
 - Binary/Boolean
 - Categorical
 - Continuous
 - Hybrid
- Relations
 - Sequential
 - Temporal
 - Spatial
 - Spatio-temporal
 - **Long range correlations**
 - **Graph**
- Data Quality
 - Data Fusion
 - Data Cleansing
 - Consistency maintenance
- Processing
 - Online/Offline processing
 - Distributed processing
 - **Analysis × Production**
 - Feature/Property searching/selection
 - Selected features detection
- Data Volume
 - Dense/**Sparse**
 - Low/High dimensions
 - Low/**Large volumes**
 - Big data
 - Internet of Things



Simple Example - Multidimensional Space ^[CBK09]

- N_1 and N_2 are regions of “normal” behavior
- Points o_1 and o_2 are anomalies
- Points in region O_3 are anomalies



Normal behavior

- **Normal distribution** ... $N(\mu, \sigma)$.
Further, it will be referred as **Gaussian** distribution
- **Normal behavior/pattern** ... it is expected, not anomalous.

Anomalies? Outliers?

Anomaly



Clouds of points (multi-dimensional)

Complex Network

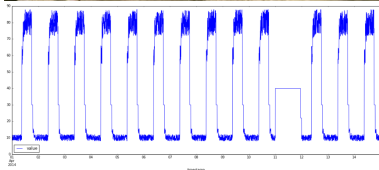
Anomalies? Outliers?

Anomaly



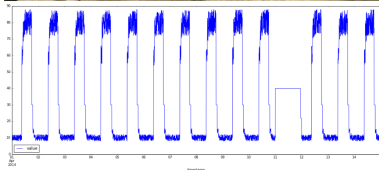
Clouds of points (multi-dimensional)

Complex Network

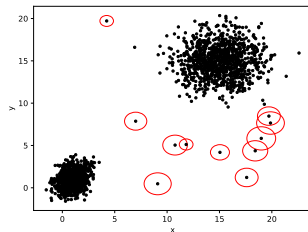


Anomalies? Outliers?

Anomaly



Clouds of points (multi-dimensional)

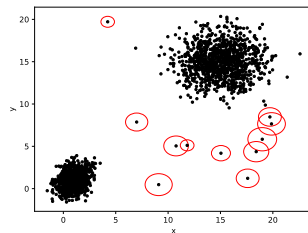


Complex Network

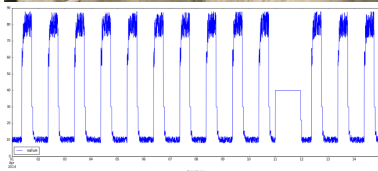
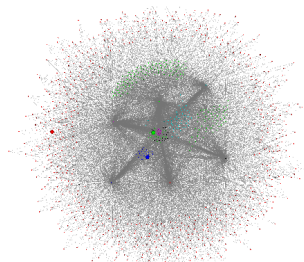
Anomalies? Outliers?

Clouds of points (multi-dimensional)

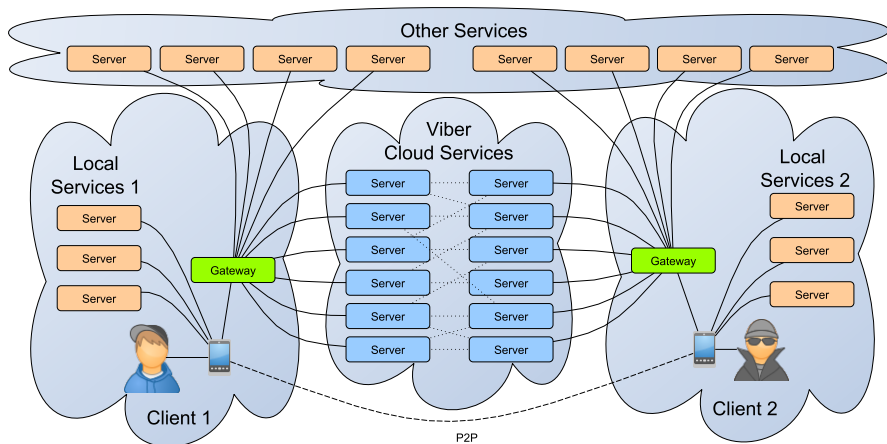
Anomaly



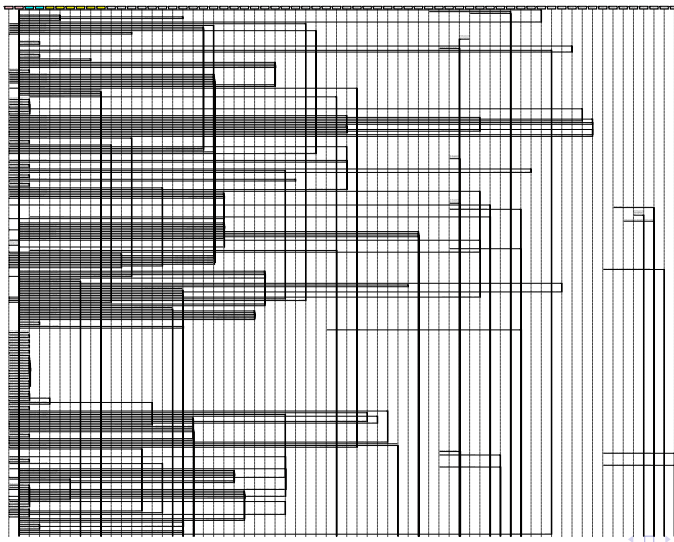
Complex Network



Exemplar (Viber) Environment ^[MBKK15]



Example Capture Characteristics - Message Sequences ^[MBKK15]

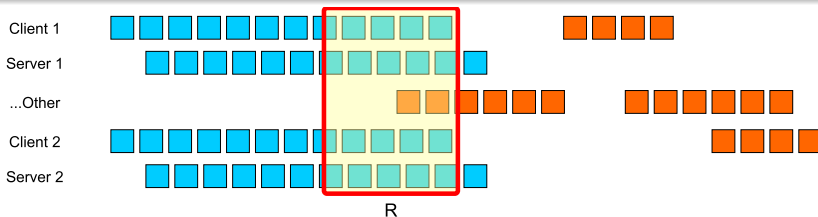


- 138882 PCAP blocks
- 1788 transport sessions
- 2 clients
- 22 viber.com servers
- 150 peers of 2 clients
- 5660 possible concurrent sessions
- How to analyze?

Concurrent Communication Detection ^[MBKK15]

Selection of IP nodes

- *viber.com* servers → viber clients → other Viber servers
- Classified based on entropy based characteristics of TCP/IP distributions

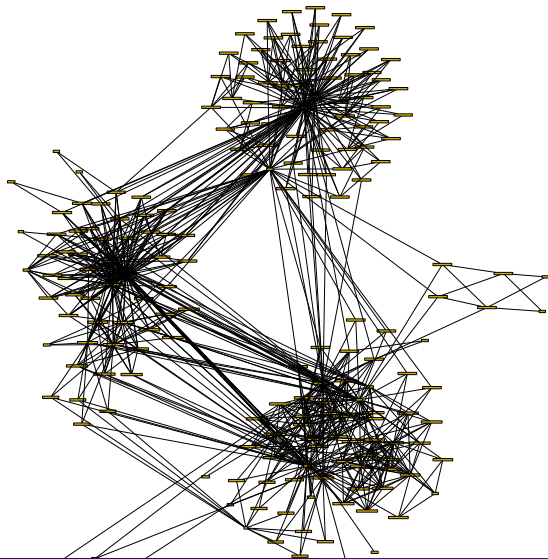


$$s(a, b) = \frac{\sum_{\forall i, j: t_a[i] - t_b[j] < R} R / (t_a[i] - t_b[j])}{\sum_{\forall i, j: t_a[i] - t_b[j] < R} 1}$$

In our experiments: $R = 50ms$, $s(a, b) > 0.001$



UDP Packet Sequence Concurrency as a Complex Network ^[MBKK15]



- Captures with two clients
- **Communities** of concurrent sessions
- Some clusters related to only one client
- Interesting clusters consist of nodes of **both** clients



Hierarchical or Partitional Clustering [XW05, EK10, vHS09]

• **Partitional clustering** (CZ rozkladové shlukování)

- the objects are divided into non-overlapping, unnested, clusters
- Given a set of input patterns $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$
- It attempts to seek a K -partition of $\mathbf{X}$, $C = \{C_1, \dots, C_K\}$, ($K \leq N$) so that
 - ① $C_i \neq \emptyset, i = 1, \dots, K$;
 - ② $\bigcup_{i=1}^K C_i = \mathbf{X}$;
 - ③ $C_i \cap C_j = \emptyset; i, j = 1, \dots, K$ and $i \neq j$.

• **Hierarchical clustering** (CZ hierarchické shlukování)

- the clusters are nested and can be displayed as a tree
- It attempts to construct a tree-like nested structure partition of $\mathbf{X}$, $H = \{H_1, \dots, H_Q\}$, ($Q \leq N$) so that

$$C_i \in H_m, C_j \in H_\ell, \text{ and } m > \ell \implies C_i \subset C_j \text{ or } C_i \cap C_j = \emptyset \quad (4)$$

$$\text{for all } i, j \neq i, m, l = 1, \dots, Q \quad (5)$$

- The tree-like partition can visualizes as a **dendrogram**

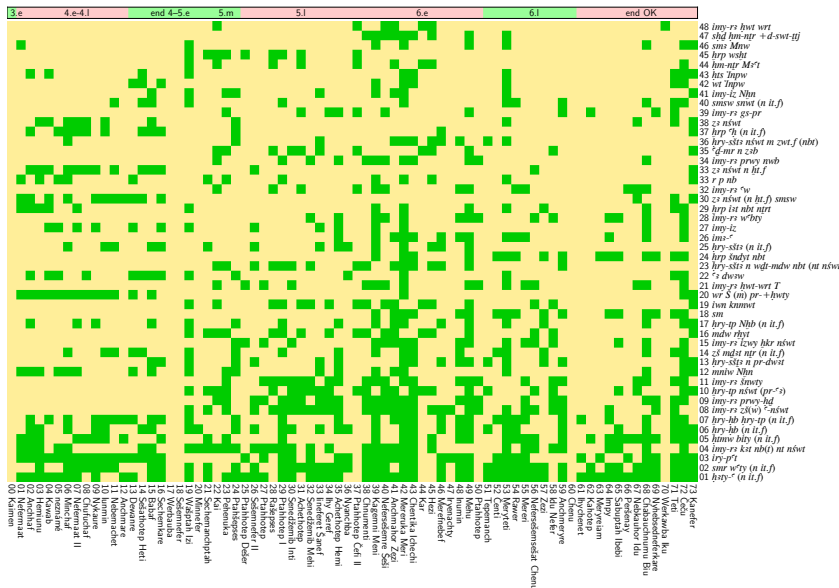


Input Data - The Old Kingdom of Egypt ^[MD15]



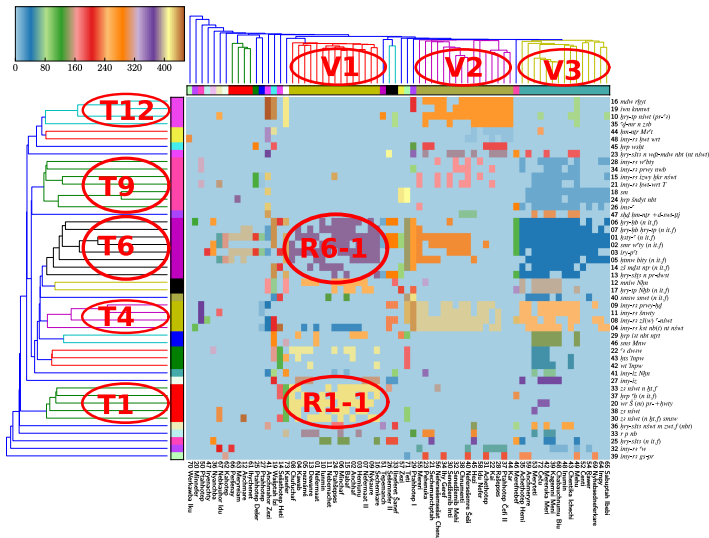
- Continuous ... tomb dimensions
- Categorical ... titles
- Binary, boolean ... titles
- Multivariate ... people, titles, tombs
- Temporal ... dynasties, king reigns
- Spatio-temporal ... location of tombs in time



Titles of Viziers ^[DMBC17]

Titles of Viziers - Jaccard, Single Linkage Clustering

[DMBC17, JD88]



Clustering [XW05, EK10]

- The steps of **cluster analysis**:
 - ① **Feature selection and extraction** selects distinguishing features from a set of candidates.
 - ② **Clustering algorithm design or selection** in which a proximity measure, a criterion function and an algorithm is determined.
 - ③ **Cluster validation** is performed to provide the users with a degree of confidence that the clustering results make sense.
 - ④ **Results interpretation** in which experts in the relevant fields interpret the data partition.
- The assigned **membership** of the nodes in the resulting clusters:
 - **Disjoint clusters** ... each node is a member of exactly one cluster.
 - **Overlapping clusters** ... a node may be a member of more than one cluster.
 - **Fuzzy clustering** methods assign a membership weight between 0 and 1 to each node such that
1 means absolute membership,
0 means a non-member.



K-Means Clustering [For65, Mac67, Har75, HW79, Llo06, XW05, EK10]

- Proposed by Lloyd in 1957 and published in 1982 ^[Llo06] and by Forgy in 1965 ^[For65]
- The term “k-means” was first used by MacQueen in 1967 ^[Mac67]
- A partitional clustering
- Given a set of observation $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$
- The aim to minimize the within-cluster sum of squares (WCSS)

$$WCSS = \underset{C}{\operatorname{argmin}} \sum_{i=1}^K |C_i| \operatorname{Var}(C_i) = \underset{C}{\operatorname{argmin}} \sum_{i=1}^K \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

- The mean $\boldsymbol{\mu}_i$ of the data points within the cluster C_i :

$$\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

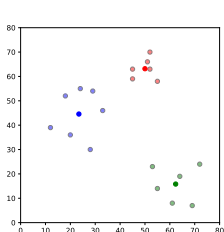


K-Means Algorithm

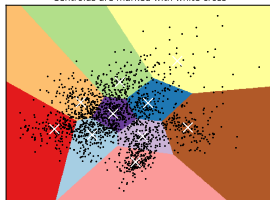
[HW79, EK10]

K -means Algorithm Basic Structure

- 1: **Input:** $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$
- 2: **Input:** K ▷ a given number of clusters
- 3: **Output:** K clusters
- 4: **select** K points as the initial centroids
- 5: **repeat**
- 6: **assign** each data point to its closest centroid and form clusters
- 7: **compute** the centroid for each cluster
- 8: **until** centroids do not change significantly



K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross

[PVG⁺11]

Outline

1 Node Influence

- Node Roles
- Hubs and Authorities

2 Network Clustering

- Multidimensional Clustering
- **Clustering Coefficients**
- Graph Laplacian



Summary of Approaches ^[New06, Weh13, CRTV07, HK13]

- The **density** of graph is the proportion of present lines to the maximum possible number of lines.
- **Clustering coefficient** is a measure of the degree to which nodes in a graph tend to cluster together

Global clustering coefficient ^[HK13]

the ratio of the total number of triangles to the total number of connected triplets.

$$C_g = \frac{2 \sum_{i=1}^N \ell_i}{\sum_{i=1}^N d_i(d_i - 1)}$$

- **Modularity** ... is - up to a normalization constant - the number of edges within communities c minus those for a **null model**
 - *"A good division of a network into communities is not merely one in which there are few edges between communities; it is one in which there are fewer than expected edges between communities".*



Clustering, Triplets and Triangles ^[CRTV07, HK13]

Clustering coefficient

- a measure of the degree to which nodes in a graph tend to cluster together
- a group of nodes with a relatively high density of ties
- ℓ_i ... the number of edges between neighbors of v_i

A triplet

- **open triangle** ... three nodes connected by two undirected ties
- of nodes (v_j, v_i, v_k) is called **connected** if v_j is connected to v_i , v_i is connected to v_k , and $j < k$. Formally, if $A_{ji} = 1$, $A_{ik} = 1$, and $j < k$.

A triangle

- **closed triangle** ... a triplet connected by three undirected ties
- is a connected triplet (v_j, v_i, v_k) in which v_j and v_k are connected. Formally, if $A_{jk} = 1$.

Clustering, Triplets and Triangles ^[CRTV07, HK13]

Clustering coefficient

- a measure of the degree to which nodes in a graph tend to cluster together
- a group of nodes with a relatively high density of ties
- ℓ_i ... the number of edges between neighbors of v_i

A triplet

- **open triangle** ... three nodes connected by two undirected ties
- of nodes (v_j, v_i, v_k) is called **connected** if v_j is connected to v_i , v_i is connected to v_k , and $j < k$. Formally, if $A_{ji} = 1$, $A_{ik} = 1$, and $j < k$.

A triangle

- **closed triangle** ... a triplet connected by three undirected ties
- is a connected triplet (v_j, v_i, v_k) in which v_j and v_k are connected. Formally, if $A_{jk} = 1$.

Clustering, Triplets and Triangles ^[CRTV07, HK13]

Clustering coefficient

- a measure of the degree to which nodes in a graph tend to cluster together
- a group of nodes with a relatively high density of ties
- ℓ_i ... the number of edges between neighbors of v_i

A triplet

- **open triangle** ... three nodes connected by two undirected ties
- of nodes (v_j, v_i, v_k) is called **connected** if v_j is connected to v_i , v_i is connected to v_k , and $j < k$. Formally, if $A_{ji} = 1$, $A_{ik} = 1$, and $j < k$.

A triangle

- **closed triangle** ... a triplet connected by three undirected ties
- is a connected triplet (v_j, v_i, v_k) in which v_j and v_k are connected. Formally, if $A_{jk} = 1$.

Numbers of Triplets and Triangles ^[CRTV07, HK13]

For a specific node v_i

- a triplet ... if $j < k$ and $A_{ji}A_{ik} = 1$
- a triangle ... if $j < k$ and $A_{ji}A_{ik}A_{jk} = 1$
- a number of connected triplets

$$N_3(i) = \sum_{j < k} A_{ji}A_{ik} = d_i(d_i - 1)/2$$

- a number of choices how edges incident to v_i can be combined.
- a number triangles

$$N_{\Delta}(i) = \sum_{j < k} A_{ji}A_{ik}A_{jk}$$

- the number of edges between neighbors of v_i

Clustering coefficient ^[CRTV07, HK13]

Transitivity, Transitivity
Ratio ^[CRTV07]

$$C = \frac{3N_{\Delta}}{N_3}$$

Clustering Coefficient,
Local Clustering
Coefficient ^[CRTV07, HK13]

$$C_i = \frac{N_{\Delta}(i)}{N_3(i)} = \frac{2\ell_i}{d_i(d_i - 1)}$$

Network average clustering
coefficient ^[HK13]

$$C_{\ell} = 1/n \sum_{i=1}^N C_i$$

Global clustering
coefficient ^[HK13]

the ratio of the total number
of triangles to the total
number of connected triplets.

$$C_g = \frac{2 \sum_{i=1}^N \ell_i}{\sum_{i=1}^N d_i(d_i - 1)}$$



Clustering coefficient ^[CRTV07, HK13]

Transitivity, Transitivity
Ratio ^[CRTV07]

$$C = \frac{3N_{\Delta}}{N_3}$$

Clustering Coefficient,
Local Clustering
Coefficient ^[CRTV07, HK13]

$$C_i = \frac{N_{\Delta}(i)}{N_3(i)} = \frac{2\ell_i}{d_i(d_i - 1)}$$

Network average clustering
coefficient ^[HK13]

$$C_{\ell} = 1/n \sum_{i=1}^N C_i$$

Global clustering
coefficient ^[HK13]

the ratio of the total number
of triangles to the total
number of connected triplets.

$$C_g = \frac{2 \sum_{i=1}^N \ell_i}{\sum_{i=1}^N d_i(d_i - 1)}$$



Clustering coefficient ^[CRTV07, HK13]

Transitivity, Transitivity
Ratio ^[CRTV07]

$$C = \frac{3N_{\Delta}}{N_3}$$

Clustering Coefficient,
Local Clustering
Coefficient ^[CRTV07, HK13]

$$C_i = \frac{N_{\Delta}(i)}{N_3(i)} = \frac{2\ell_i}{d_i(d_i - 1)}$$

Network average clustering
coefficient ^[HK13]

$$C_{\ell} = 1/n \sum_{i=1}^N C_i$$

Global clustering
coefficient ^[HK13]

the ratio of the total number
of triangles to the total
number of connected triplets.

$$C_g = \frac{2 \sum_{i=1}^N \ell_i}{\sum_{i=1}^N d_i(d_i - 1)}$$



Clustering coefficient ^[CRTV07, HK13]

Transitivity, Transitivity
Ratio ^[CRTV07]

$$C = \frac{3N_{\Delta}}{N_3}$$

Clustering Coefficient,
Local Clustering
Coefficient ^[CRTV07, HK13]

$$C_i = \frac{N_{\Delta}(i)}{N_3(i)} = \frac{2\ell_i}{d_i(d_i - 1)}$$

Network average clustering
coefficient ^[HK13]

$$C_{\ell} = 1/n \sum_{i=1}^N C_i$$

Global clustering
coefficient ^[HK13]

the ratio of the total number
of triangles to the total
number of connected triplets.

$$C_g = \frac{2 \sum_{i=1}^N \ell_i}{\sum_{i=1}^N d_i(d_i - 1)}$$



Outline

1 Node Influence

- Node Roles
- Hubs and Authorities

2 Network Clustering

- Multidimensional Clustering
- Clustering Coefficients
- Graph Laplacian



Diffusion Equation ^[Cra75]

- The **diffusion equation** is a partial differential equation.
- In physics, it describes the behavior of the *collective motion* of *micro-particles* in a material resulting from the random movement of each micro-particle.

$$\frac{\partial \phi(\mathbf{r}, t)}{\partial t} = \nabla \cdot [D(\phi, \mathbf{r}) \nabla \phi(\mathbf{r}, t)]$$

- where $\phi(\mathbf{r}, t)$ is the density of the diffusing material
- at location $\mathbf{r}$ and time t and
- $D(\phi, \mathbf{r})$ is the collective diffusion coefficient for density ϕ at location $\mathbf{r}$; and
- ∇ represent the vector differential operator del (nabla).
- If D is constant, then the equation reduces to the linear differential equation (the **heat equation**), ∇^2 is the Laplacian operator:

$$\frac{\partial \phi(\mathbf{r}, t)}{\partial t} = D \nabla^2 \phi(\mathbf{r}, t) \quad \Delta = \nabla^2 = \sum_{k=1}^n \frac{\partial^2}{\partial x_k^2}$$



Diffusion on Networks ^[New10]

- **Diffusion process** as a simple model of spread across a network
 - the spread of an idea,
 - the spread of a disease/virus
- An amount ψ_i of some commodity or substance at vertex i .
- The commodity flows from vertex i to an adjacent one j at a rate $C(\psi_j - \psi_i)$
- where C is a constant called the **diffusion constant**.
- The rate at which ψ_i is changing is given by

$$\frac{d\psi_i}{dt} = C \sum_j A_{ij}(\psi_j - \psi_i)$$

- The equation works for both undirected and directed networks.
- Let us assume an undirected and simple network.



Diffusion Matrix Form Derivation ^[New10]

- Network diffusion equation

$$\frac{d\psi_i}{dt} = C \sum_j A_{ij} (\psi_j - \psi_i)$$

- Splitting, rewriting, merging

$$\begin{aligned} \frac{d\psi_i}{dt} &= C \sum_j A_{ij} \psi_j - C \psi_i \sum_j A_{ij} \\ &= C \sum_j A_{ij} \psi_j - C \psi_i k_i \\ &= C \sum_j (A_{ij} - \delta_{ij} k_i) \psi_j \end{aligned}$$

- where $k_i = \sum_j A_{ij}$ is the degree of vertex i and
- δ_{ij} is the Kronecker delta.



Diffusion Matrix Form ^[New10]

- Network diffusion equation

$$\frac{d\psi_i}{dt} = C \sum_j (A_{ij} - \delta_{ij} k_i) \psi_j$$

- In matrix form

$$\frac{d\psi}{dt} = C(\mathbf{A} - \mathbf{D})\psi$$

- where ψ is the vector whose components are amounts ψ_i ,
- $\mathbf{A}$ is the adjacency matrix, and
- $\mathbf{D}$ is the diagonal matrix with the vertex degrees along the diagonal

$$\mathbf{D} = \begin{pmatrix} k_1 & 0 & 0 & \cdots \\ 0 & k_2 & 0 & \cdots \\ 0 & 0 & k_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$



Graph Laplacian ^[New10, EK10]

- Network diffusion equation

$$\frac{d\psi}{dt} = C(\mathbf{A} - \mathbf{D})\psi$$

- In matrix form
- where ψ is the vector whose components are amounts ψ_i ,
- $\mathbf{A}$ is the adjacency matrix, and
- The **graph Laplacian** is the real symmetric matrix

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

$$L_{ij} = \begin{cases} k_i & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and there is an edge between vertices } j \text{ and } i, \\ 0 & \text{otherwise} \end{cases}$$

$$L_{ij} = \delta_{ij}k_i - A_{ij}$$



Diffusion Equation Solution ^[New10]

- Network diffusion equation

$$\frac{d\psi}{dt} = C(\mathbf{A} - \mathbf{D})\psi \quad \Rightarrow \quad \frac{d\psi}{dt} = -C\mathbf{L}\psi$$

- Assuming the vector ψ as a linear combination of the Laplacian eigenvectors $\mathbf{v}_i$, i.e. $\mathbf{L}\mathbf{v}_i = \lambda_i\mathbf{v}_i$ and $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ for $i \neq j$ (real, sym $\mathbf{L}$)

$$\psi(t) = \sum_i a_i(t) \mathbf{v}_i$$

- with the coefficients $a_i(t)$ varying over time.
- By the substitution into the diffusion equation and the dot product with $\mathbf{v}_j$

$$\sum_i \left(\frac{da_i}{dt} + C\lambda_i a_i \right) \mathbf{v}_i = 0 \quad \xrightarrow{\cdot \mathbf{v}_j} \quad \frac{da_j}{dt} + C\lambda_j a_j = 0$$

- The solution

$$a_i(t) = a_i(0)e^{-C\lambda_i t}$$



Eigenvalues of the Graph Laplacian ^[New10]

- Let $\mathbf{B}$ be the edge incidence matrix (i.e. edges $\times$ vertices)
 - If $i \neq j$ then $\sum_k B_{ki} B_{kj}$ is -1 if there is an edge between vertices i and j , and zero otherwise.
 - If $i = j$ then $\sum_k B_{ki}^2$ has a term $+1$ for every edge connected to vertex i so that the sum is equal to the degree k_i of vertex i .

$$L_{ij} = \sum_k B_{ki} B_{kj} \quad \Rightarrow \quad \mathbf{L} = \mathbf{B}^T \mathbf{B}$$

- Let $\mathbf{v}_i$ be an eigenvector of $\mathbf{L}$ with eigenvalue λ_i , i.e. $\mathbf{L} \mathbf{v}_i = \lambda_i \mathbf{v}_i$.

$$\mathbf{v}_i^T \mathbf{B}^T \mathbf{B} \mathbf{v}_i = \mathbf{v}_i^T \mathbf{L} \mathbf{v}_i = \lambda_i \mathbf{v}_i^T \mathbf{v}_i = \lambda_i \quad \text{as } \mathbf{v}_i^T \mathbf{v}_i = 1$$

- It is just an inner product of a real vector ($\mathbf{B} \mathbf{v}_i$) with itself.
 - $\Rightarrow$ It is a sum of squares $\Rightarrow$ the sum $\geq 0 \Rightarrow \lambda_i \geq 0$.
 - The solution of the diff. equation contains only decaying exponentials.
 - The solution tends to an equilibrium value as $t \rightarrow \infty$.
- $\mathbf{L} \cdot \mathbf{1} = \mathbf{0} \Leftrightarrow \sum_j L_{ij} \times 1 = \sum_j (\delta_{ij} k_i - A_{ij}) = k_i - \sum_j A_{ij} = k_i - k_i = 0$
 - $\mathbf{1}$ is always an eigenvector of $\mathbf{L}$ with the smallest eigenvalue $\lambda_1 = 0$
 - $\Rightarrow \mathbf{L}$ is singular, the Laplacian has no inverse.



Eigenvalues of the Graph Laplacian ^[New10]

- Let $\mathbf{B}$ be the edge incidence matrix (i.e. edges $\times$ vertices)
 - If $i \neq j$ then $\sum_k B_{ki} B_{kj}$ is -1 if there is an edge between vertices i and j , and zero otherwise.
 - If $i = j$ then $\sum_k B_{ki}^2$ has a term $+1$ for every edge connected to vertex i so that the sum is equal to the degree k_i of vertex i .

$$L_{ij} = \sum_k B_{ki} B_{kj} \quad \Rightarrow \quad \mathbf{L} = \mathbf{B}^T \mathbf{B}$$

- Let $\mathbf{v}_i$ be an eigenvector of $\mathbf{L}$ with eigenvalue λ_i , i.e. $\mathbf{L}\mathbf{v}_i = \lambda_i \mathbf{v}_i$.

$$\mathbf{v}_i^T \mathbf{B}^T \mathbf{B} \mathbf{v}_i = \mathbf{v}_i^T \mathbf{L} \mathbf{v}_i = \lambda_i \mathbf{v}_i^T \mathbf{v}_i = \lambda_i \quad \text{as } \mathbf{v}_i^T \mathbf{v}_i = 1$$

- It is just an inner product of a real vector ($\mathbf{B}\mathbf{v}_i$) with itself.
 - $\Rightarrow$ It is a sum of squares $\Rightarrow$ the sum $\geq 0 \Rightarrow \lambda_i \geq 0$.
 - The solution of the diff. equation contains only decaying exponentials.
 - The solution tends to an equilibrium value as $t \rightarrow \infty$.
- $\mathbf{L} \cdot \mathbf{1} = \mathbf{0} \Leftrightarrow \sum_j L_{ij} \times 1 = \sum_j (\delta_{ij} k_i - A_{ij}) = k_i - \sum_j A_{ij} = k_i - k_i = 0$
 - $\mathbf{1}$ is always an eigenvector of $\mathbf{L}$ with the smallest eigenvalue $\lambda_1 = 0$
 - $\Rightarrow \mathbf{L}$ is singular, the Laplacian has no inverse.



Algebraic Connectivity ^[New10]

$$\mathbf{L} = \left(\begin{array}{cc|cccc} & & 0 & 0 & 0 & 0 \\ & L_1 & 0 & 0 & 0 & 0 \\ & & 0 & 0 & 0 & 0 \\ 0 & 0 & & L_2 & 0 & 0 \\ 0 & 0 & & & 0 & 0 \\ 0 & 0 & 0 & 0 & & \ddots \\ 0 & 0 & 0 & 0 & & \ddots \end{array} \right)$$

$$\mathbf{v} = (\underbrace{1, 1, 1, \dots}_{n_1 \text{ ones}}, \underbrace{0, 0, 0, \dots}_{\text{zeros}})$$

- Suppose we have a network that is divided up into c different components of sizes $n_1, n_2, \dots, n_c$.
- $\implies$ at least c eigenvectors with eigenvalue zero
- The number of zero eigenvalues is always exactly equal to the number of components.
- The second eigenvalue λ_2 is non-zero if and only if the network is connected, it is called the **algebraic connectivity**.



Summary

- Node roles
- Hubs and Authorities (HITS)
- Data Clustering
 - Introduction and examples
 - K-means clustering
- Clustering, triplets, and triangles
- Diffusion Equation
 - Graph Laplacian



Competencies

- What are the basic roles of nodes?
- How is it possible to assess a role of a given nodes?
- Provide definitions of authorities and hubs.
- How are the hub and authority centralities defined?
- What is the goal of clustering?
- What are the two fundamental approaches to data clustering?
- What are the typical steps of a cluster analysis?
- What are the basic forms of node memberships in clusters?
- Describe k-means clustering.
- Define a triplet and triangle.
- Describe a diffusion equation.
- What is the graph Laplacian?
- Name basic properties of the graph Laplacian eigenvalues?



References I

- [Agg17] Charu C. Aggarwal. *Outlier Analysis*. Springer, second edition, 2017.
- [CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.
- [Cra75] J. Crank. *The mathematics of diffusion*. Clarendon Press, second edition, 1975.
- [CRTV07] L. D. F. Costa, F. A. Rodrigues, G. Traverso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56:167–242, January 2007.
- [DMBC17] Veronika Dulíková, Radek Mařík, Miroslav Barta, and Matej Cibulá. HMM model vývoje a trendů správy země v období Staré říše. In *16. ročník konference Počítačová podpora v archeologii, Písek CZ, 29. - 31. května 2017*. Katedra archeologie Západočeské univerzity v Plzni, CZ, 2017.
- [EK10] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets. Reasoning About a Highly Connected World*. Cambridge University Press, July 2010.
- [For65] E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21(3):768–769, 1965.
- [Har75] J.A. Hartigan. *Clustering Algorithms*. John Wiley & Sons Inc., New York, 1975.
- [HK13] Stephen J. Hardiman and Liran Katzir. Estimating clustering coefficients and size of social networks via random walk. In *Proceedings of the 22Nd International Conference on World Wide Web, WWW '13*, pages 539–550, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [HW79] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.
- [JD88] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.



References II

- [Kle98] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '98, pages 668–677, Philadelphia, PA, USA, 1998. Society for Industrial and Applied Mathematics.
- [Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, September 1999.
- [Llo06] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, September 2006.
- [Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.
- [MBKK15] Radek Mařík, Pavel Bezpalec, Jan Kučerák, and Lukáš Kencl. Revealing viber communication patterns to assess protocol vulnerability. In *2015 International Conference on Computing and Network Communications (CoCoNet)*. Leonia, NJ 07605: EDAS Conference Services, pages 502–510, 2015.
- [MD15] Radek Mařík and Veronika Dulíková. *Mathematical Formalization of Society Complexity*, chapter Povaha změny: Bezpečnost, rizika a stav dnešní civilizace, pages 98–129. Praha Vyšehrad, ISBN 978-80-7429-641-3, 2015. (in Czech).
- [New06] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- [New10] M. Newman. *Networks: an introduction*. Oxford University Press, Inc., 2010.
- [PVG⁺11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.



References III

- [STE07] [Jerry Scripps, Pang-Ning Tan, and Abdol-Hossein Esfahanian](#). Node roles and community structure in networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 26–35, New York, NY, USA, 2007. ACM.
- [vHS09] [Hana Řezanková, Dušan Húsek, and Václav Snášel](#). *Shluková analýza dat*. Professional Publishing, Praha, second edition, 2009.
- [Weh13] [Stefan Wehrli](#). Social network analysis, lecture notes, December 2013.
- [XW05] [Rui Xu and D. Wunsch](#). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, May 2005.

