

Software Engineering Department

Braude College

Capstone Project Phase A – 61998

**Temporal identifying of influential nodes**

**in complex citation networks**

**25-1-R-18**

Students:

Oren Baranovsky – 311170450 [oren.barnovsky@e.braude.ac.il](mailto:oren.barnovsky@e.braude.ac.il)

Anna Garmash – 337889125 [Anna.Garmash@e.braude.ac.il](mailto:Anna.Garmash@e.braude.ac.il)

Supervisors: Dr. Renata Avros

Prof. Zeev Volkovich

Git: <https://github.com/orenbara/influential_nodes>

**Table of contents**

[**Abstract** 4](#_Toc190289093)

[**1.** **Introduction** 4](#_Toc190289094)

[**2.** **Background and related work** 5](#_Toc190289095)

[**2.1** **Node Centrality** 5](#_Toc190289096)

[**2.1.1** **Degree centrality** 6](#_Toc190289097)

[**2.1.2** **Betweenness centrality** 7](#_Toc190289098)

[**2.1.3** **Closeness centrality** 7](#_Toc190289099)

[**2.1.4** **Eigenvector centrality** 8](#_Toc190289100)

[**2.2** **Gravity Models** 10](#_Toc190289101)

[**2.2.1** **Classic Gravity Model** 11](#_Toc190289102)

[**2.2.2** **Local Gravity Model** 11](#_Toc190289103)

[**2.2.3** **Clustering Gravity Model** 12](#_Toc190289104)

[**2.2.4** **Network Embedding Gravity Model** 12](#_Toc190289105)

[**2.2.5** **Random Walk-Based Gravity Model** 12](#_Toc190289106)

[**2.2.5.1** **Random Walk** 12](#_Toc190289107)

[**2.2.5.2** **Gravity model based on random walk** 13](#_Toc190289108)

[**2.2.6** **Effective Distance Gravity Model** 13](#_Toc190289109)

[**2.2.6.1** **Effective Distance of distant nodes** 14](#_Toc190289110)

[**2.2.6.2** **Distance of nodes with multiple path options** 14](#_Toc190289111)

[**2.2.6.3** **Effective Distance is asymmetric** 15](#_Toc190289112)

[**2.2.6.4** **Effective Distance Gravity Model (EFFG)** 15](#_Toc190289113)

[**2.2.7** **Weighted Gravity Model** 15](#_Toc190289114)

[**3.** **Research Process** 17](#_Toc190289115)

[**3.1** **Node importance in citation networks** 17](#_Toc190289116)

[**3.2** **Expected behavior** 17](#_Toc190289117)

[**3.3** **The process** 18](#_Toc190289118)

[**3.3.1** **Data preprocessing for citation network analysis** 18](#_Toc190289119)

[**3.3.2** **Computing the importance scores** 19](#_Toc190289120)

[**3.3.3** **Storing score results** 20](#_Toc190289121)

[**4.** **Result analysis** 21](#_Toc190289122)

[**4.1** **Measures** 21](#_Toc190289123)

[**4.1.1** **Standard deviation** 21](#_Toc190289124)

[**4.1.2** **Pearson correlation coefficient** 21](#_Toc190289125)

[**4.1.3** **Spearman's rank correlation coefficient** 22](#_Toc190289126)

[**4.1.4** **Kendall’s tau** 24](#_Toc190289127)

[**4.1.5** **SIR - Susceptible-Infected-Recovered** 26](#_Toc190289128)

[**4.2** **Approach to result analysis** 27](#_Toc190289129)

[**5.** **Related Diagram** 29](#_Toc190289130)

[**6.** **Testing plan** 30](#_Toc190289131)

[**7.** **References** 31](#_Toc190289132)

# **Abstract**

This paper explores different methods that make use of Newton’s gravity model to calculate the importance scores of nodes in a growing citation network. The importance scores are required to identify special behaviors of documents like exponential growth in interest, clustering around subjects, and increase in authority. Three different models were chosen for the research. Local Gravity is the first model, utilizing the degree, shortest path, and truncation radius it provides scores for all the nodes in the graph. The second model utilizes Effective Distance which was introduced in the disease spreading area and derived from the probability of disease spread in the network. The third model uses eigenvalues to provide higher scores of importance for nodes that are connected to neighbors with a high degree of influence. For each document, the results are stored and analyzed using tools like graph visualization, Spearman’s rank, and Kendall rank correlation coefficient.

# **Introduction**

Complex networks are a quickly growing field of study. Researches in this area come from fields like mathematics, computer science, physics, biology, epidemiology, sociology, and more [[1](#bookmark=id.30j0zll)]. Thus, the results are both theoretically important and practical in these areas. Citation networks are a type of complex network where nodes represent academic papers and edges signify citation relationships between them. These networks are necessary for studying how knowledge spreads across various scientific fields.

The identification of the most influential nodes in a complex citation network is essential for revealing scientific trends. In particular, recently published works with fast-growing citations may point to a new and relevant idea. Moreover, this approach can be applied in order to improve educational resources. Understanding which works have the greatest impact can help both lecturers and students select key materials for reading or learning.

The existing research articles that study the influential nodes often apply their models to social, collaborative, transportation, technology, infrastructure, communication, or epidemic-spreading networks [[2](#bookmark=id.1fob9te)]. However, the use of these methods on citation networks still remains unexplored. Traditional methods mostly work with static networks and do not consider the temporal dynamics of node influence. Nevertheless, in real-world networks node influence changes over time [[3](#bookmark=id.3znysh7)].

This research’s goal is to apply existing models for studying influential nodes to the citation graph and examine how effective these methods are in analyzing scientific paper networks. In addition, a major focus is to investigate the dynamics of node influence over time. The impact of an article is not static in citation networks: new works can become more popular while older ones may lose relevance. Considering the temporal aspect helps us understand how and why the significance of nodes changes.

# **Background and related work**

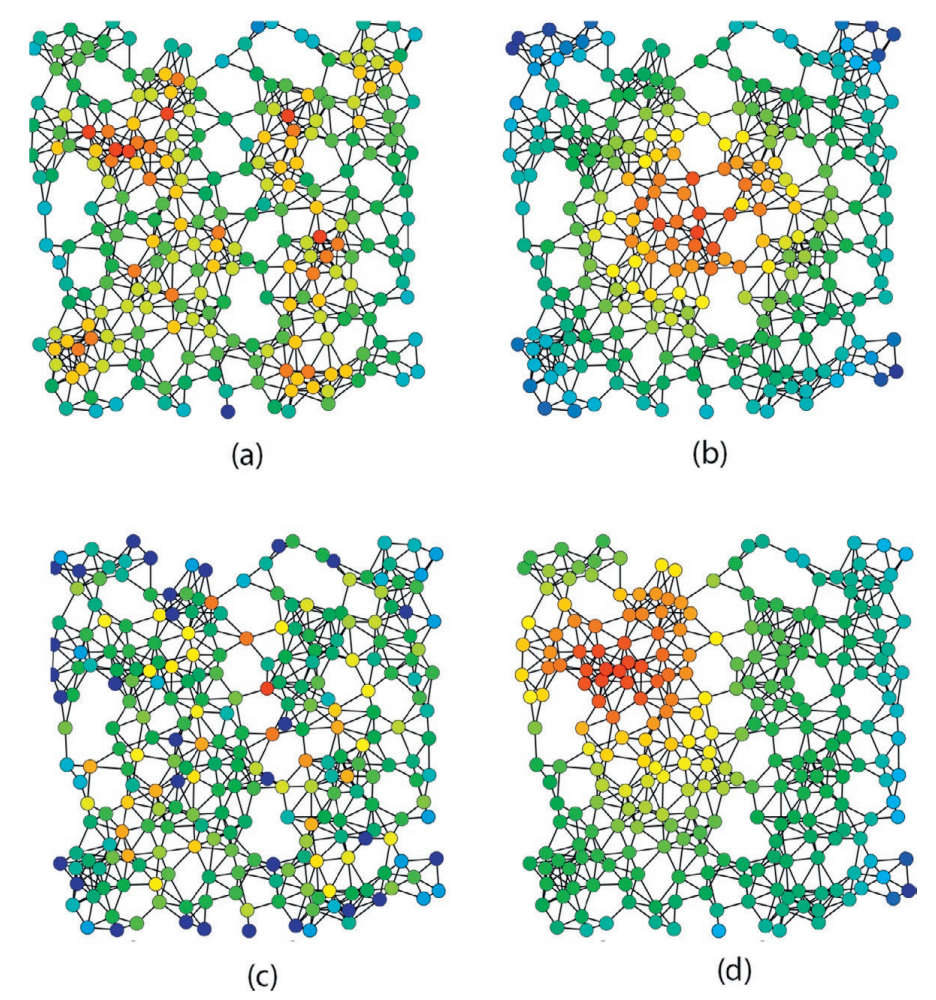
A lot of research has already been applied to the topic of identifying influential or important nodes in a complex network. In this research, we apply Gravity models to citation networks to find and observe the influential nodes. In the following section, we will discuss the existing Centrality methods and Gravity models.

In a directed graph , V represents the set of nodes and E represents the edges. The number of nodes in the graph is n, where . A is the adjacency matrix of G, , where = 1 if there is an edge between node i and node j.

## **Node Centrality**

A node's centrality in a graph measures its importance or influence within the network. Different centrality metrics capture different aspects of importance based on **local**, and **global** information. Local information might include properties of a node like in-degree, out-degree, k-shell value, H-index, and more [[gpt1](#bookmark=id.uialdzys1qnc)]. The global information refers to properties that describe the structure such as network size, density, diameter, average path length, degree distribution.

Centrality measuring might include local, global, or hybrid information in the calculation. Next are some known centrality measurement methods for a node.



***Figure 1.*** *Red nodes are more central and blue nodes are less central. Version (a) is degree centrality, (b) uses closeness centrality, (c) shows betweenness centrality, and (d) is eigenvector centrality. This visualization is adapted from Claudio Rocchini.* ***[***[***20***](#bookmark=id.c0vijewkqg53)***]***

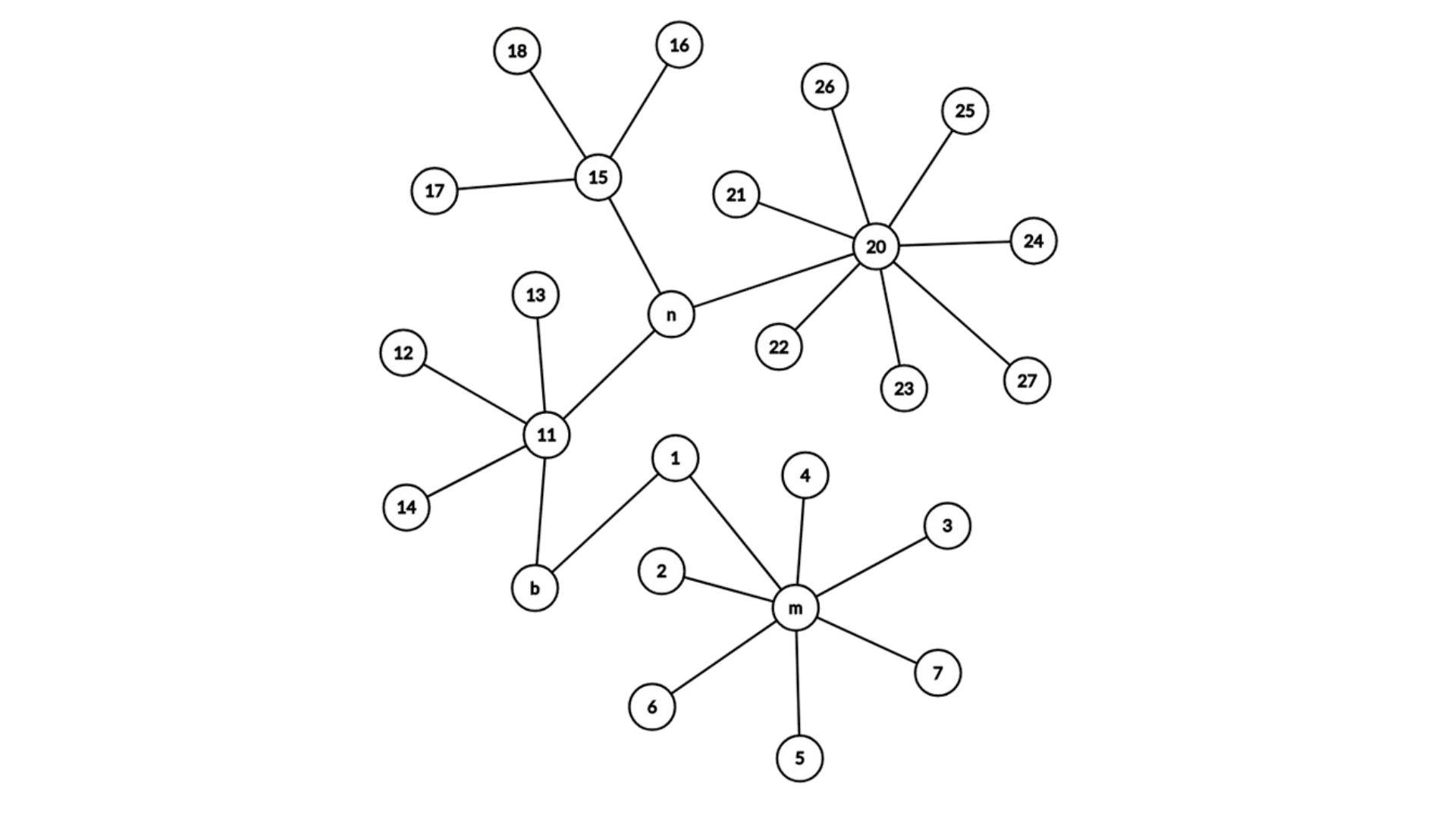
### **Degree centrality**

Nodes with a high degree of centrality are influential because they interact with many others [[4](#bookmark=id.2hbmphakt1xv)].



***Figure 2.*** *Green and yellow nodes are more central and blue nodes are less central.* [[21](#bookmark=id.whhc627xgtyn)]

Degree centrality is a straightforward and efficient metric. Due to its simplicity, in some cases, it fails to identify the most influential nodes.



***Figure 3.***Although n has a lower DC than m, in some networks n is considered much more important than m.

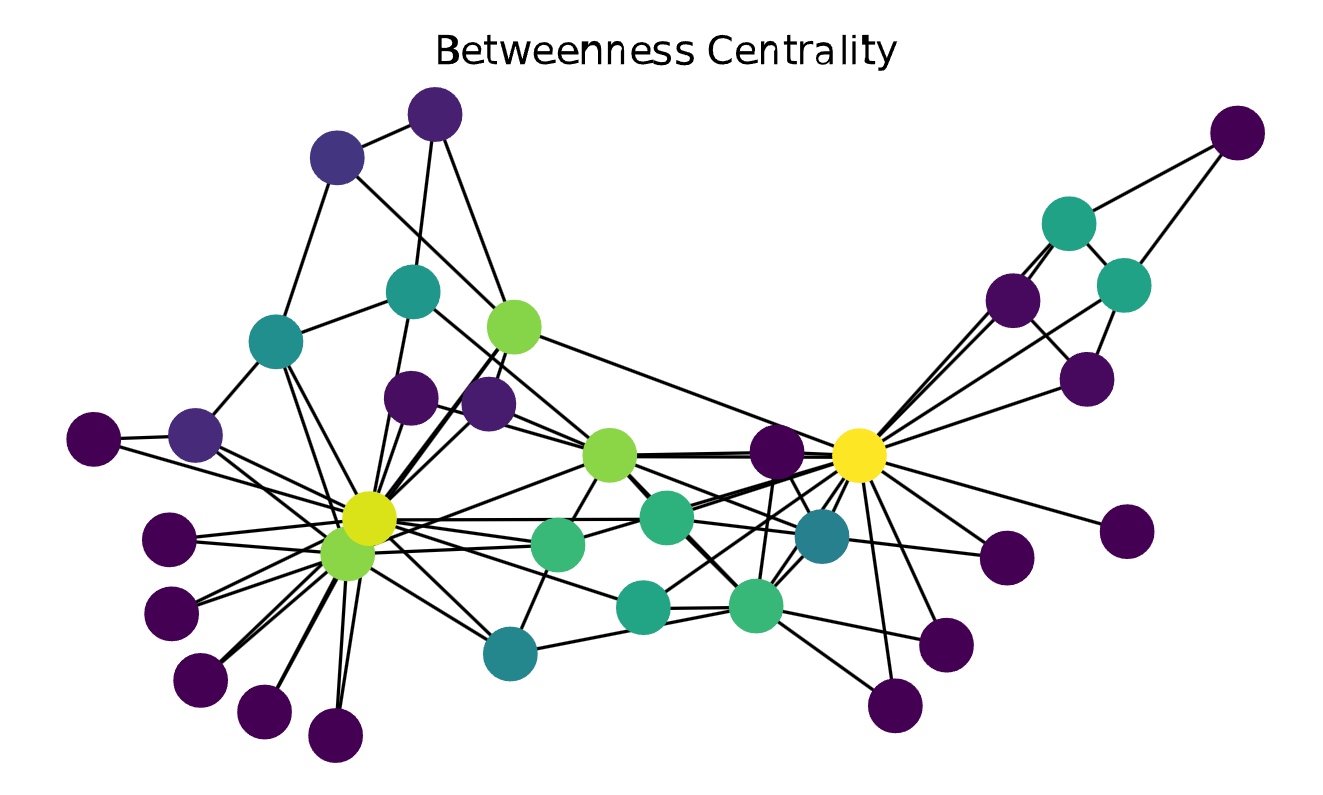
### **Betweenness centrality**

Measures the importance of a node by the number of shortest paths through it [[5](#bookmark=id.uuzboljcretz)].

where represents the number of shortest paths from node *j* to node *k*, and includes only the paths which contain *i*.

Nodes with high betweenness centrality control communication between different parts of the network.

Betweenness is, in some sense, a measure of the influence of a node over the information spread through the network or the expected load of a node in a transportation network [[4](#bookmark=id.2hbmphakt1xv)].



***Figure 4.*** *Green and yellow nodes are more central and blue nodes are less central.* ***[***[***21***](#bookmark=id.whhc627xgtyn)***]***

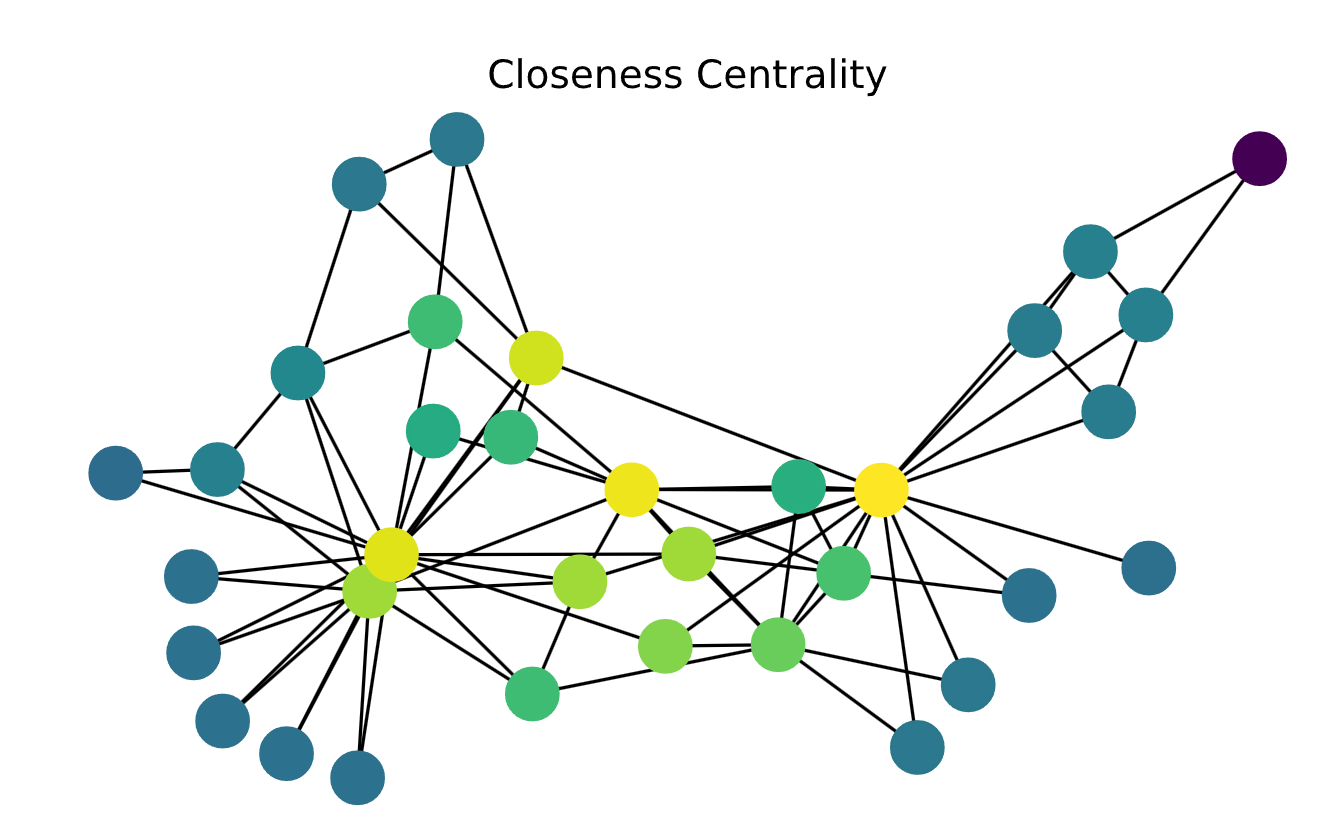
### **Closeness centrality**

Evaluate centrality using the reciprocal of the sum of shortest paths [[4](#bookmark=id.2hbmphakt1xv)].

where denotes the shortest length of path between node *i* and node *j*.

[[gpt1](#bookmark=id.uialdzys1qnc)] Nodes with high closeness centrality can quickly interact with all other nodes in the network.

Closeness can be considered a measure of how long it will spread information from a given node to other reachable nodes in the network [[4](#bookmark=id.2hbmphakt1xv)].



***Figure 5.*** *Green and yellow nodes are more central and blue nodes are less central.* ***[***[***21***](#bookmark=id.whhc627xgtyn)***]***

### **Eigenvector centrality**

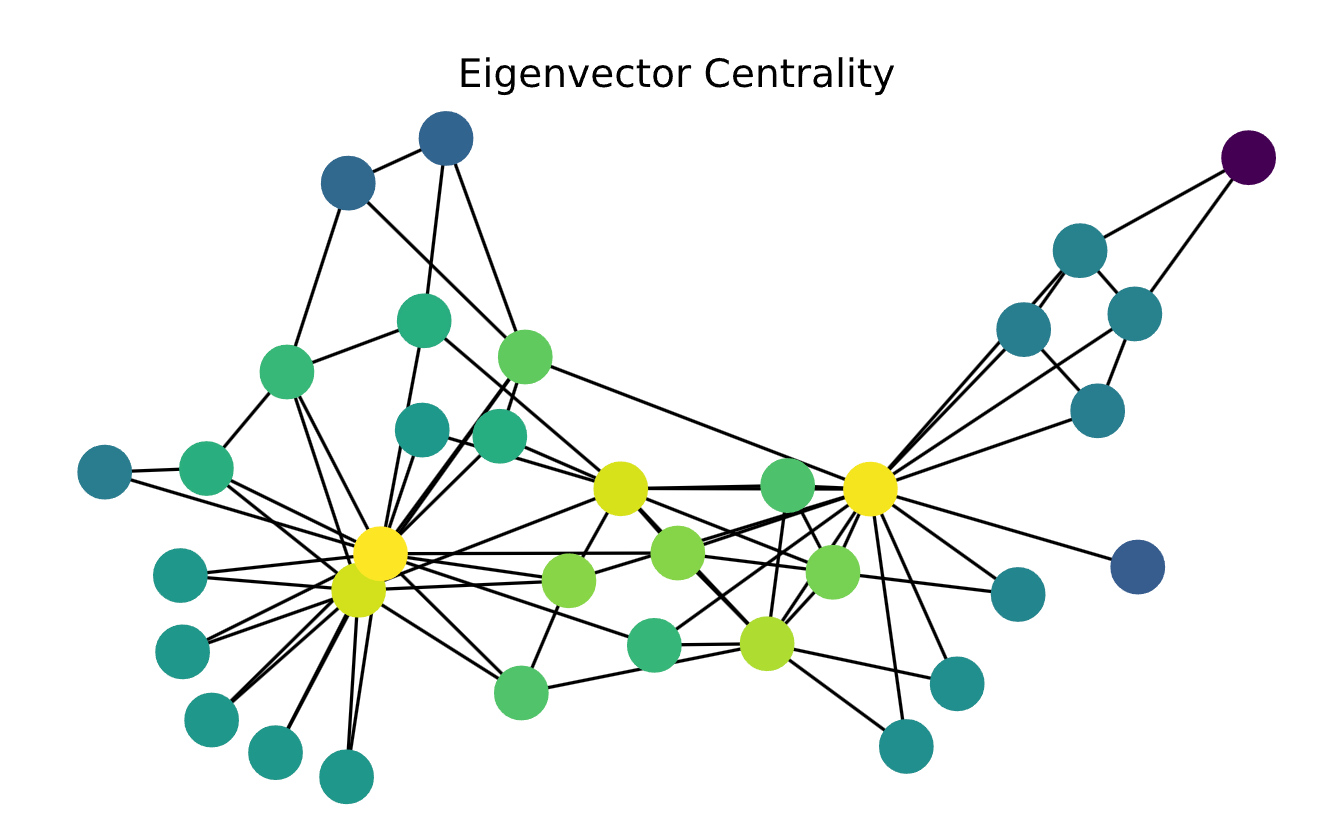
Assigns higher importance to nodes that are connected to other influential nodes [[6](#bookmark=id.om6383162bv4)].

.

The largest eigenvalue of A is represented by and is the value of *j-*th entry of the eigenvector corresponding to .

Nodes with high eigenvector centrality are connected to other high-centrality nodes, making them globally influential.

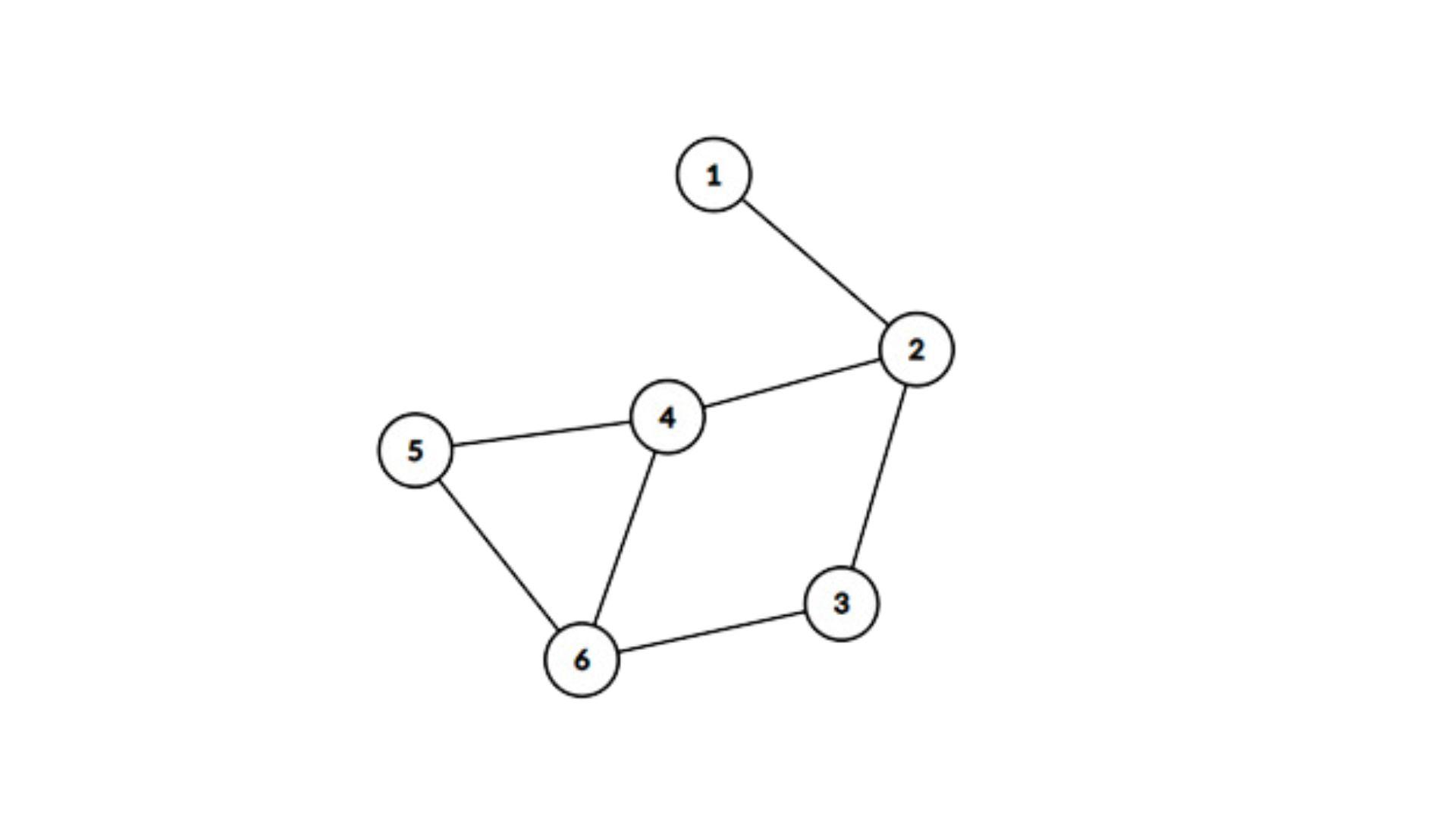
A node may have a high degree score but a relatively low Eigenvector Centrality score if any connections are with similarly low-scored nodes. Also, a node may have a high betweenness score but a low Eigenvector centrality score because it might be distant from the power sources in the network.

****

***Figure 6.*** *Green and yellow nodes are more central and blue nodes are less central.* ***[***[21](#bookmark=id.whhc627xgtyn)***]***

Quantifying the node's centrality which represents connectivity to important neighbors based on the eigenvector's corresponding value is unintuitive. Understanding the following algorithm of node centrality calculation is the key to understanding the correlation between eigenvectors and the centrality discussed. Since important nodes gain their power from neighbors with high degrees, the importance of all nodes will be calculated in an iterative process. In each iteration, the Centrality Degrees of all the neighbors of node i will be summed and this value will replace the importance score from the last iteration. As the iterations progress the score grows and thus at the end, the scores are normalized so that the largest score is 1. Following this algorithm, and after final normalization, the vector representation of the scores converges to the eigenvector corresponding to the highest eigenvalue of the adjacency matrix.

The following example illustrates this approach:



***Figure 7.*** *Example of a simple graph.*

The highest eigenvalue of the adjacency matrix of the graph in the example above corresponds to the eigenvector: [0.31, 0.79, 0.69, 1.00, 0.78, 0.97].

When applying the algorithm with four iterations before normalizing the scores the results of the scores are as follows:

***Table 1.*** *Nodes’ scores by algorithm.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 |
| 1 | 1 | 1 | 3 | 6 | 17 |
| 2 | 1 | 3 | 6 | 17 | 38 |
| 3 | 1 | 2 | 6 | 13 | 37 |
| 4 | 1 | 3 | 8 | 19 | 52 |
| 5 | 1 | 2 | 6 | 15 | 39 |
| 6 | 1 | 3 | 7 | 20 | 47 |

The 4th score vector is [17,38,37,52,39,47] and the normalized vector is [0.32, 0.73, 0.71, 1.00, 0.75, 0.90].

The result is similar to the eigenvector:

Eigenvector: [0.31, 0.79, 0.69, 1.00, 0.78, 0.97]

Algorithm result: [0.32, 0.73, 0.71, 1.00, 0.75, 0.90]

Using algebraic Spectral decomposition, it could be proven that this algorithm converges to the eigenvector.

## **Gravity Models**

Gravity models introduce an innovative approach to measure the influence or importance of a node. The intuition for these models is inspired by Newton's law of universal gravitation, where the gravitational force between objects depends on their mass and distance. The formula is:

where *F* is the gravitational force, *G* is the gravitational constant, and are the masses of the object, and *r* is the distance between their centers.

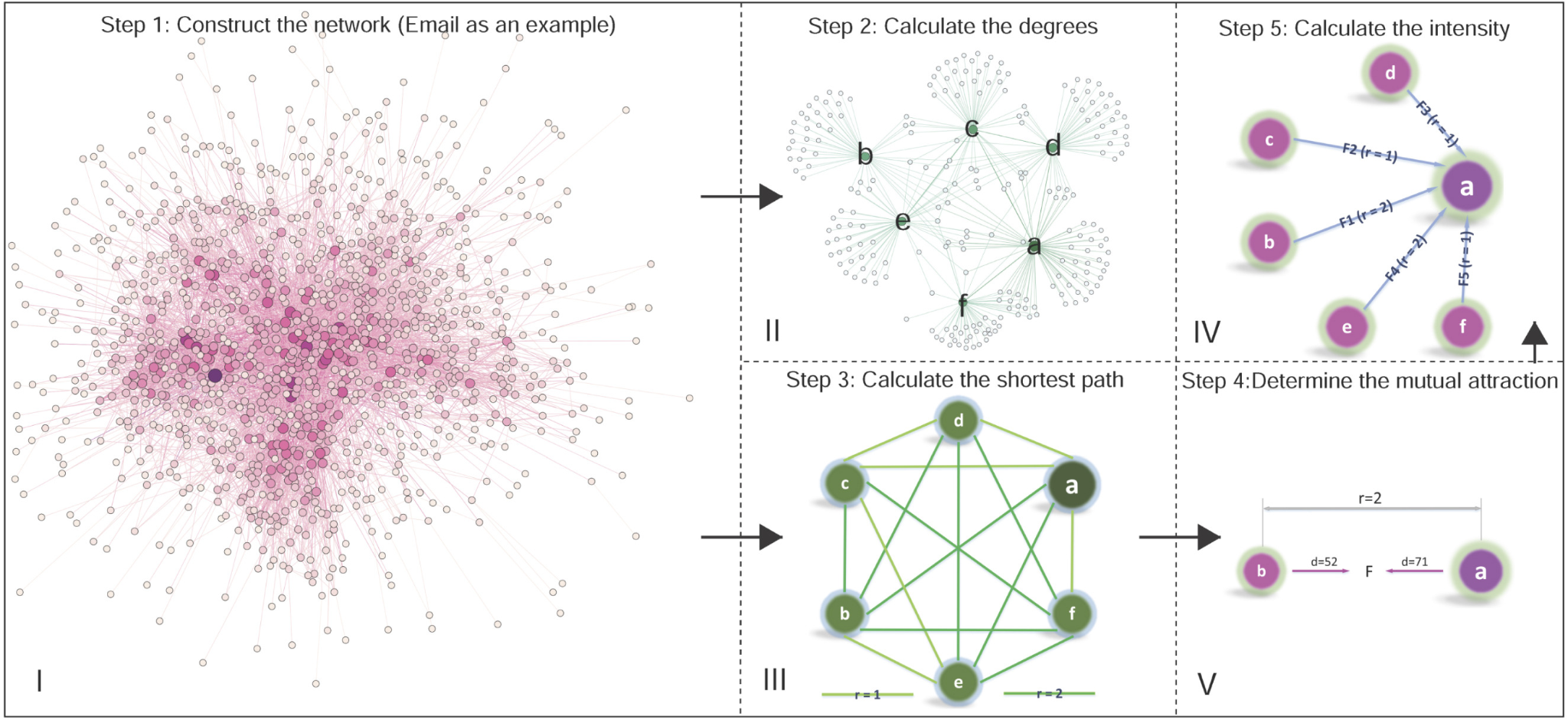
In the context of gravity models beyond physics, each model redefines “mass”, “distance” and the “gravity coefficient” based on the elements it considers important. This concept was first introduced at [[7](#bookmark=id.8cax1a8hpz4a)]. The intuition behind gravity models lies in their ability to capture interaction potential. Entities with larger “mass” apply a stronger pull on each other, while greater separation decreases interaction strength. This balance between mass and distance reflects a universal principle of how connections are influenced by size and proximity.

Gravity models are particularly effective for analyzing networks because they can capture the balance between node importance (“mass”) and their relationships (“distance”). The flexibility of redefining parameters makes gravity models adaptable to different domains. This makes them ideal for understanding complex networks because they provide insights into both local and global network behavior.

### **Classic Gravity Model**

A classic version of Gravity calculation for a node in the graph.

where is the degree of node *i* and is the shortest distance between nodes *i* and *j*.



***Figure 8.*** *The flow chart of the Classic Gravity Model.* ***[***[***7***](#bookmark=id.8cax1a8hpz4a)***]***

### **Local Gravity Model**

When nodes are located far away their gravity force is weak. The Local Gravity Model suggested by Li et al. [[8](#bookmark=id.9j7aw8qydahq)] allows for simplifying the calculations by limiting the amount of distance calculation only to nodes of predefined distance R.

,

where R is the truncation radius and is often set to half of the average length of the shortest path.

### **Clustering Gravity Model**

Li et al. [[10](#bookmark=id.arlp9yqzf56m)] proposed to redefine the mass: use instead of node degree . This model allows quantifying the effect of nodes with large clustering (neighbor connectivity).

where is a configurable parameter, that denotes the number of edges between neighbors of node *i*.

### **Network Embedding Gravity Model**

This model was suggested in [[11](#bookmark=id.cde17gjgkw8w)] and modifies the distance by embedding the nodes into low-dimensional vectors in Euclidean space.

### **Random Walk-Based Gravity Model**

Most gravity models have an expensive time complexity of finding the shortest distance between all nodes, according to the complexity of the Dijkstra algorithm. The authors of the paper [[15](#bookmark=id.bv30n8d3wy7p)] propose replacing the shortest distance with the node’s position in a random walk, reducing the complexity.

#### **Random Walk**

Random walk on a graph is a walk strategy where it starts from the node v and moves to one of its neighbors x with probability . The outcome of the process is *l*-length walks of each node where the probability of visiting x from node v can be defined as:

where *Z* is a normalization constant and – the unnormalized jumping probability between nodes v and x.

Jumping probability indicates how likely it is for the walker to jump from one node to another. Two tunable parameters are used to control how the walk explores neighboring nodes.

* *p* – return parameter. It measures the possibility of revisiting the node during the walk. A higher value reduces the probability of revisiting a node.
* *q* - in-out parameter. It controls the strategy of the walk. The strategies could be, for example, Breadth-First Search (BFS) or Depth-First Search (DFS).

In addition, two more important parameters must be considered:

* γ – the number of random walks performed for each node
* *l* – the length of each random walk

Choosing the correct value for *γ* determines the measure of graph coverage during random walks. As *γ* increases, the graph is explored more fully, meaning that more random walks are performed for each node. A similar principle applies to the parameter *l*. Increasing the walk length allows the exploration of more distant parts of the graph, providing a more detailed understanding of its structure and relationships.

#### **Gravity model based on random walk**

The input data for the algorithm are pre-generated random walk samples. For each node, starting from the second position in each random walk sample, the algorithm calculates the gravity value based on:

with

where is the degree of *j*-th (position) node in the *m*-th walk. is the number of walks for each node and *l* is the length of the walk.

It is important to note that the value *j*-1 is used because it means the distance (the number of steps) from the source node to the *j*-th node. This is the replacement for the shortest distance, as mentioned earlier.

### **Effective Distance Gravity Model**

In **[**[**14**](#bookmark=id.9vcp27insdbo)**] Effective Distance** was introduced to replace the geographical distance of the paths connecting nodes of cities to predict the time in which a disease will spread to a target city. In **[**[**14**](#bookmark=id.9vcp27insdbo)**]** Effective Distance revealed complex hidden patterns of the network which provided more accurate information for the task given compared to the physical geographical distance.

In [[16](#bookmark=id.4j2y2ehsa73z)] **Effective Distance** was merged into the Gravity Model as the distance between two nodes, replacing the Euclidean distance and shortest path distance. **Effective Distance** [[16](#bookmark=id.4j2y2ehsa73z)] is a distance calculated from the probability, which reveals the hidden pattern geometry of the network.

The effective distance from node m to node n which are directly connected is defined as:

where is the value of the effective distance from node *m* to node *n* if they are directly connected. is the probability of moving from node *m* to node *n*.

where is the degree of node m and is the element corresponding to the edge between m and n (1 if there is an edge between *m* and *n*).

Effective Distance is asymmetric, meaning that .

#### **Effective Distance of distant nodes**

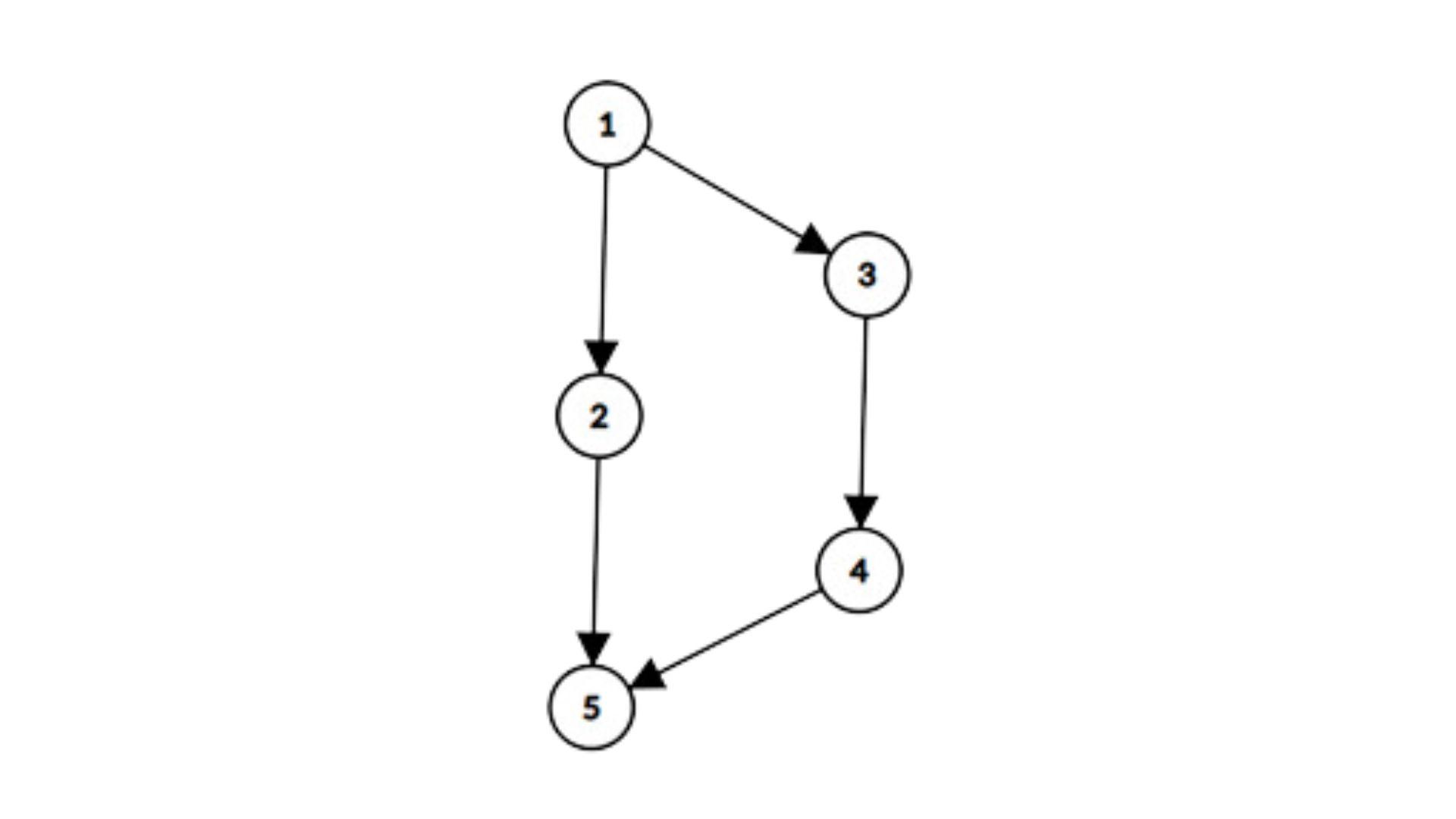
The effective distance of nodes that are not directly connected can be obtained transitively.



***Figure 9.*** *Example of indirect connection.*

#### **Distance of nodes with multiple path options**

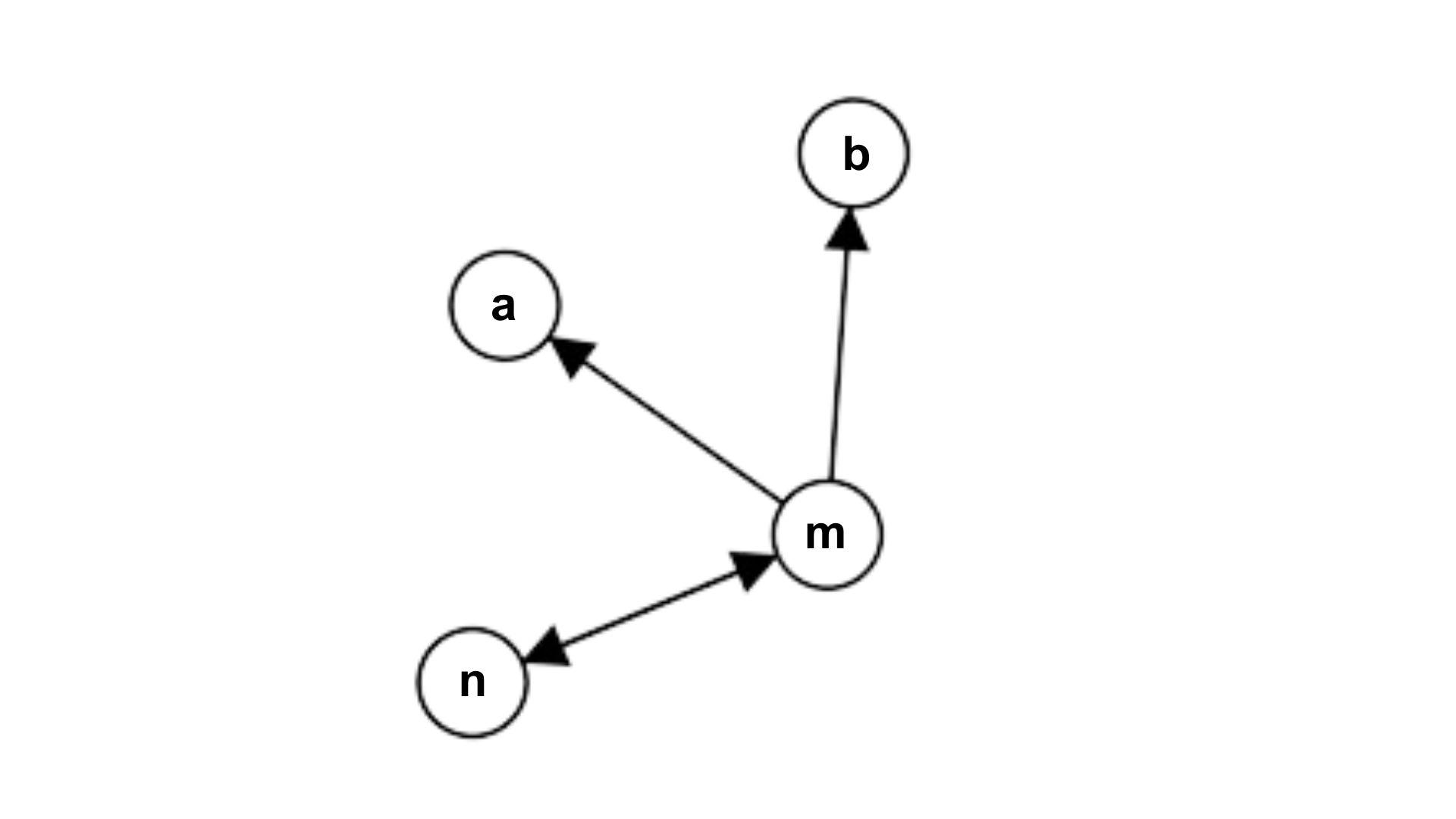
The shortest path is chosen if there are multiple paths from node m to point n.

****

***Figure 10.*** *Example of multiple paths.*

If there are no weights then the path { } will be chosen.

#### **Effective Distance is asymmetric**

****

***Figure 11.*** *Example of asymmetric relationships.*

[[16](#bookmark=id.4j2y2ehsa73z)] To identify important nodes in real-world networks it is insufficient to consider only local and global node information, moreover, it is not enough to consider static and dynamic information. Often there is hidden structural information about the interaction between nodes which is important for measuring the importance of nodes.

#### **Effective Distance Gravity Model (EFFG)**

This model was introduced in [[16](#bookmark=id.4j2y2ehsa73z)] to calculate the importance of nodes in real-world networks.

where and are the degree of node *i* and node *j* respectively

is the effective distance from node i to node j.

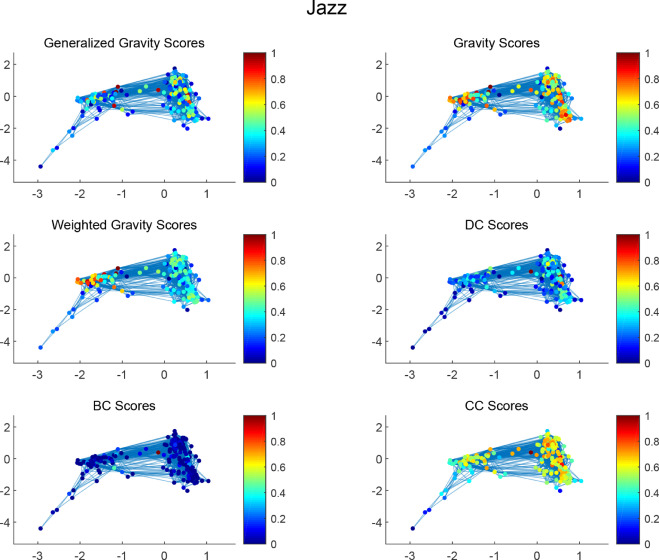
denotes the interaction score between node *i* and node *j*, which might be considered as the gravity force in which the nodes pull each other.

### **Weighted Gravity Model**

This method was suggested in [[9](#bookmark=id.8kayppm6uamh)] and utilizes the idea of eigenvector centrality and Truncation radius.

The weighted gravity model also called GMM: generalized mechanical model is proposed that uses global information and local information [[9](#bookmark=id.8kayppm6uamh)]. If the degree of one node is larger (local information), the shorter the distance and the greater the weight (global information) with other nodes, the influence of the node is greater.

where is the largest eigenvalue of . X is the eigenvector. is the *i*-th value of X.



***Figure 12.*** *Gravity models applied to the famous Jass network.* ***[***[***19***](#bookmark=id.jc1h7efu6676)***]***

# **Research Process**

## **Node importance in citation networks**

In this work, the effectiveness of Gravity Models in citation networks was researched. A citation graph (or citation network), is a directed graph that describes the citations within a collection of documents. Each vertex in the graph represents a document in the collection, and each edge is directed from one document toward another that it cites (or vice versa depending on the specific implementation).

During the research, lots of methods to identify influential nodes were studied and Gravity Based Models were chosen for the task of identifying different kinds of important nodes. Out of all the different types of important nodes that exist in network-based data, three types were chosen:

* **High Citation Nodes**: documents that are cited a lot.
* **Bridge Nodes**: documents that provide connectivity between clusters.
* **Authoritative Nodes**: documents that are cited by other important documents which make them more authoritative.

The idea is to identify these three types of important nodes using different gravity models and track the behavior of the importance scores throughout the timestamps for each of these nodes.

The following tools were chosen for implementation:

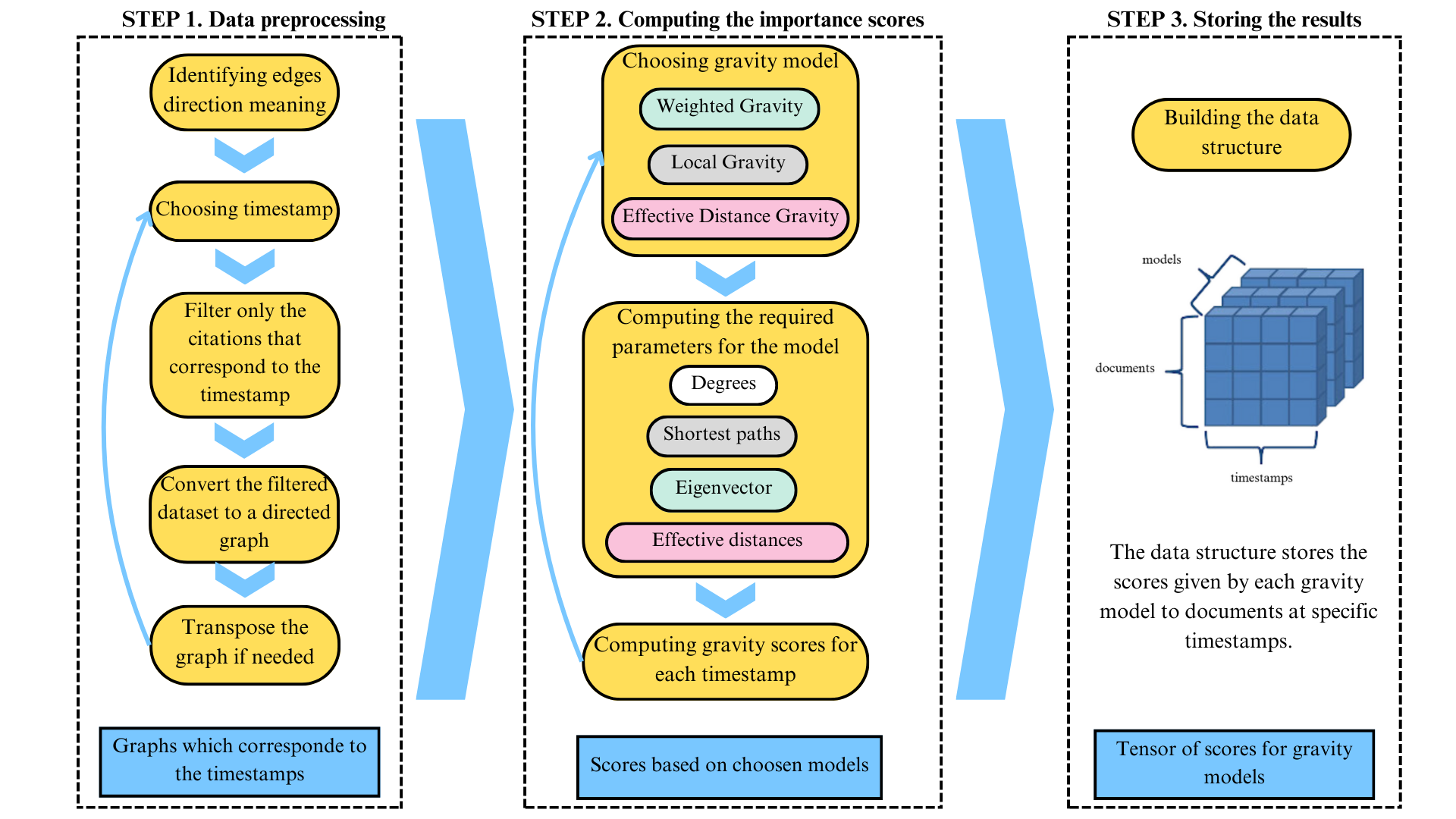
* Python and R languages and libraries like Pandas for data set management.
* Python Networkx and Geometry (PYG) for graph requirements.
* Python plotting libraries like Pyplot for visual result analysis.

## **Expected behavior**

Since the **Local Gravity Model** defines the degree as its mass and the shortest path as the distance, it is a good candidate for identifying the **High Citation Nodes.**

The **Effective Distance Gravity Model** reveals the hidden pattern geometry of the network, this information is used to identify **Bridge Nodes.**

**Authoritative Nodes** are identified by the **Weighted Gravity Model** since it makes use of the eigenvectors to provide a higher score of importance for nodes that connect to other important nodes.



***Figure 13.*** *Project flow.*

## **The process**

### **Data preprocessing for citation network analysis**

The research is about dynamic (temporal) citation networks, so data that consists of document citations is required. Best matches should include additional information about the publication year to capture the important changes through the timestamps.

Data preprocessing is essential to achieve an accurate analysis of citation networks. The preprocessing includes the generation of a directed graph based on the citation information file. There is high importance on the direction of the edges - it should be clear which document cites and which is cited for the models to run correctly.

Suppose the data set includes timestamps, the dataset is divided into subsets based on timestamps. These timestamps can present intervals such as years, decades, or custom time frames suited to the characteristics of the specific dataset. The choice of time intervals depends on the destiny of citations and the frequency of new article publications. Dividing the datasets into subsets is necessary for studying the dynamics of node influence over time.

The process results in a directed graph that is represented by an adjacency matrix.

### **Computing the importance scores**

To identify the different types of important nodes three gravity model versions are used. Assuming the provided graph defines citation of node n by node m as a directed edge from n to m (the other way around is also acceptable but has to be accounted for in the implementation), identification of the nodes with the **highest citation count** as important and done by the **Local Gravity Model** via executing the following stages:

1. For each node *i* in the graph, the out-degree is calculated.
2. For each node, *i* and each , the shortest path between *i* and *j* and then the pull force is calculated.
3. For each node *i*, the total sum of all the pull forces corresponding to each is calculated.

Identification of **Bridge Nodes** is done by the **Effective Distance Gravity Model** executing the following stages:

From [[16](#bookmark=id.4j2y2ehsa73z)]:

1. Calculate the effective distance between all nodes
2. Calculate the interaction scores between all pairs of nodes.
3. Use cumulative sum to get the EffG centrality score of each node.

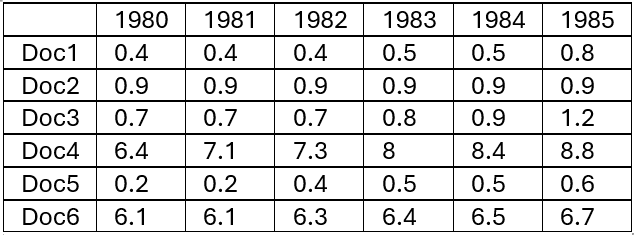
Identification of **Authoritative Nodes** is done by the **Weighted Gravity Model:**

From [[9](#bookmark=id.8kayppm6uamh)]:

1. Construct network
2. Calculate degree:
3. Calculate the value of the largest normalized eigenvector
4. Calculate the interaction force of nodes:

### **Storing score results**

After obtaining the gravity scores for the data of the first timestamp, the results are stored for analysis in a data structure. The data structure contains score information for each document throughout the timestamps. For every subsequent timestamp, new scores are calculated and appended to the structure. The final version of the data structure captures the dynamic evolution of influence scores.



***Figure 14.*** *Scores of nodes over time, where the columns are chosen timestamps and the rows are the documents that were analyzed.*

Expected challenges:

* Large citation data sets are required with timestamps, locating the datasets with such restrictions might become an issue.
* In large networks computation might take long time, in the project some models require heavy computation like calculating shortest paths.
* The chosen gravity models might not provide high scores for the types of importance that were defined.

# **Result analysis**

## **Measures**

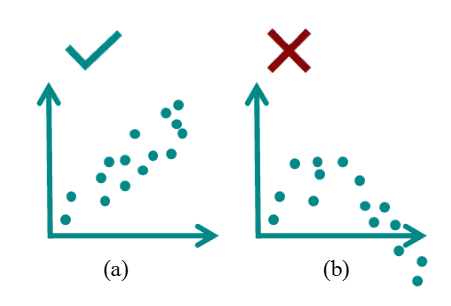
### **Standard deviation**

where is each data point in the set, is the mean and *N* is the total number of data points.

Standard deviation () [[18](#bookmark=id.h5aocy8okqh1)] is a key statistical measure that indicates how dispersed the data points are in relation to the mean. A standard deviation close to zero indicates that data points are very close to the mean, whereas a large standard deviation indicates data points are spread further away from the mean.

### **Pearson correlation coefficient**

Pearson correlation coefficient [[23](#bookmark=id.x8e9myxf7mf6)] examines the relationship between two variables, i.e. how large the linear relationship is. It measures only linear relationships, thus Pearson correlation doesn’t suit nonlinear relationships.

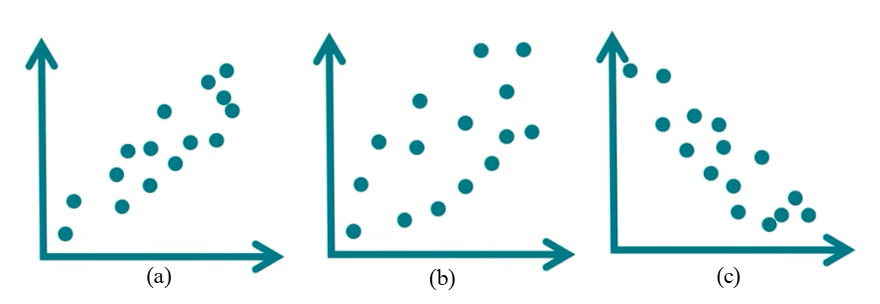


***Figure 15.*** *Example of linear (a) and nonlinear (b) relationships.* ***[***[***23***](#bookmark=id.x8e9myxf7mf6)***]***

It is the ratio between the covariance of two variables and the product of their standard deviations.

where is the Pearson correlation coefficient, *X* and *Y* are random variables, *cov* is the covariance, and the standard deviation of *X* and *Y*, is the individual value of one variable, is the individual value of another variable, and are respectively the mean values of the two variables. The result always has a value between -1 and 1.

The coefficient determines how strong the correlation is and in which direction it goes. A positive relationship or correlation occurs when high values of one variable are associated with high values of another, or when low values of one variable are associated with low values of the other. A negative correlation occurs when high values of one variable correspond to low values of the other and on the contrary.



***Figure 16.*** *Examples of positive correlation (a), no correlation (b), and negative correlation (c).* ***[***[***23***](#bookmark=id.x8e9myxf7mf6)***]***

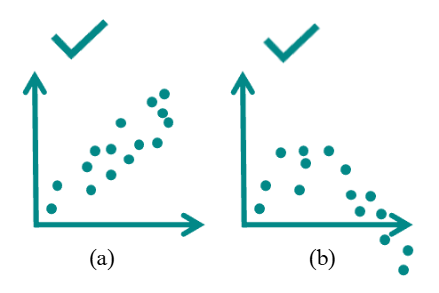
The strength of the correlation is represented in the following way [[26](#bookmark=id.811yad65fpdx)]:

***Table 2.*** *The strength of the correlation.*

|  |  |  |
| --- | --- | --- |
| **value (positive)** | **value (negative)** | **Strength of the correlation** |
| 0.0 < 0.1 | -0.1 < 0.0 | No correlation |
| 0.1 < 0.3 | -0.3 < -0.1 | Low correlation |
| 0.3 < 0.5 | -0.5 < -0.3 | Medium correlation |
| 0.5 < 0.7 | -0.7 < -0.5 | High correlation |
| 0.7 < 1 | -1 < -0.7 | Very high correlation |

### **Spearman's rank correlation coefficient**

The Spearman’s rank correlation coefficient [[24](#bookmark=id.8p8b1k31gwyl)] examines the relationship between two variables. To compute the Spearman correlation, we calculate the Pearson correlation using the ranks of the data instead of the original values. Essentially, the Spearman correlation is equivalent to the Pearson correlation, but it operates on ranks rather than raw data. Therefore this allows to measure as linear as the non-linear relationships.



***Figure 17.*** *Example of linear (a) and nonlinear (b) relationships.* ***[***[***24***](#bookmark=id.8p8b1k31gwyl)***]***

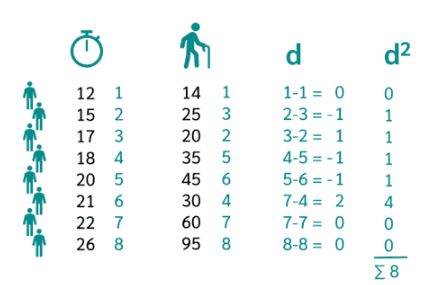
In statistics, ranking is a transformation of numerical or ordinal values into ranks based on their position in a sorted dataset. For example, if the data is represented by: 14, 25, 20, 35, 60, 30, 90, the ranks after sorting would be: 1, 3, 2, 5, 6, 4, 7.

where is Spearman’s rank correlation coefficient, is the Pearson correlation coefficient, *X* and *Y* are random variables, *R[X]* and *R[Y]* are ranks, *cov* is the covariance, and are standard deviations of rank values.

If all ranks are unique values and ranking is straightforward (i.e., there are no tied ranks), a simplified formula can be used:

where is Spearman’s rank correlation coefficient, *d* is the difference between the two ranks of each observation and *n* is the number of observations.

For example, the calculation of Spearman’s rank correlation coefficient of the reaction time of 8 computer gamers of different ages looks like this [[24](#bookmark=id.8p8b1k31gwyl)]:

****

***Figure 18.*** *Example of Spearman’s rank correlation coefficient calculation.* ***[***[***24***](#bookmark=id.8p8b1k31gwyl)***]***

where 12, 15, 17, 18, 20, 21, 22, 26 are time reactions with the ranks 1, 2, 3, 4, 5, 6, 7, 8 respectively. The numbers 14, 25, 20, 35, 45, 30, 60, 90 are the people’s age with the ranks 1, 3, 2, 5, 6, 4, 7, 8 respectively.

Like Pearson's correlation coefficient, Spearman's correlation coefficient also varies between -1 and 1. Using Spearman’s coefficient, the strength and direction of correlation can be determined exactly as with Pearson's coefficient.

### **Kendall’s tau**

Kendall’s Tau [[25](#bookmark=id.uhwhf1bzuqmw)] is a correlation coefficient that measures the relationship between two variables. It is very similar to the Spearman correlation. For the calculation of Kendall’s Tau, the two variables need only have ordinal scale levels. However, Kendall's tau should be preferred to Spearman's correlation when only a few data with many ties are available, i.e. two or more ranks are equal.

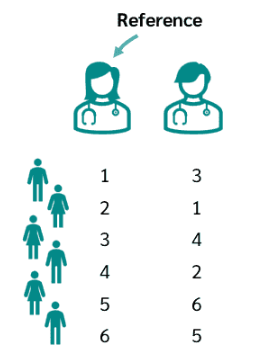
where *C* is the number of concordant pairs and *D* is the number of discordant pairs.

Observed pairs are concordant if the ranks for and move in the same direction. That means:

Observed pairs are discordant if the ranks for and move in the opposite direction. That means:

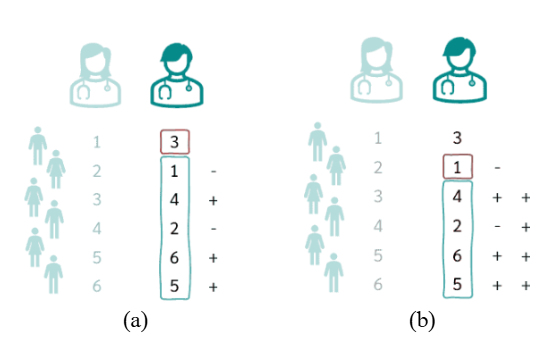
For example, the calculation of Kendall’s Tau correlation coefficient looks like this [[25](#bookmark=id.uhwhf1bzuqmw)]:

Two doctors ranked six patients based on their physical health, from best to worst. In this scenario, the female doctor is designated as the reference, and the patients are assigned ranks from 1 to 6. The sorted ranks can be compared to the rankings assigned by the second doctor. The main purpose is to calculate the correlation between the two assessments.



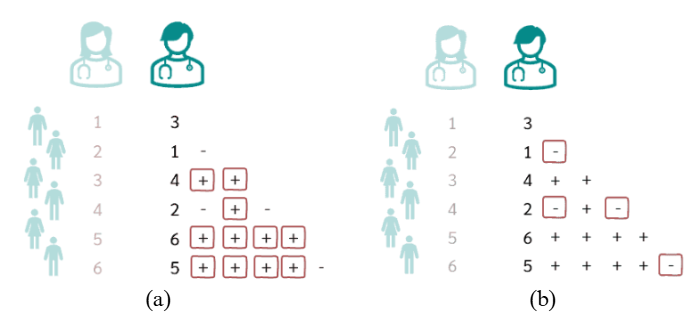
***Figure 19.*** *The example of the problem for Kendall’s Tau coefficient calculation.* ***[***[***25***](#bookmark=id.uhwhf1bzuqmw)***]***

To determine the concordant and discordant pairs, each rank compares its value to the value of all ranks below it, checking whether each is smaller or larger, like in the figure below:



***Figure 20.*** *The example of rank comparison for the first (a) and the second rank (b).* ***[***[***25***](#bookmark=id.uhwhf1bzuqmw)***]***

Thus, the number of concordant pairs is the number of “+” and the number of discordant pairs is the number of “-”. In the example, C is equal to 11, while D is equal to 4:



***Figure 21.*** *The example of concordant (a) and discordant (b) pairs calculation.* ***[***[***25***](#bookmark=id.uhwhf1bzuqmw)***]***

Hence, the Kendall’s Tau correlation coefficient in the example:

Kendall’s Tau correlation coefficient varies between -1 and 1. Using Kendall’s Tau coefficient, the strength and direction of correlation can be determined exactly as with Pearson's coefficient.

### **SIR - Susceptible-Infected-Recovered**

This is a classic epidemiological model designed to analyze the spread of infectious diseases within a population. It divides the population into three groups: susceptible (S), infected (I), and recovered (R), describing the dynamics of transitions between these states over time. Initially proposed for studying epidemics, the SIR model is also applied in other fields [[17](#bookmark=id.v10h16wwvri6)].

Adapted SIR model [[gpt2](#bookmark=id.irts9dfqs9s0)] for a citation graph can be represented as follows:

* Susceptible (S) - papers that have not yet been cited;
* Infected (I) - papers that are actively being cited;
* Recovered (R) - papers that are no longer being significantly cited.

Parameters definitions:

* *β* (Contact Rate) - the likelihood that a cited paper will “infect” other papers, causing them to cite it;
* *γ* (Recovery Rate) - the rate at which a paper becomes less influential over time.

if S(t) is the number of uncited or not-yet-influential papers at time t; I(t) is the number of actively cited papers at time t; R(t) is the number of papers that have lost influence by time t.

By analyzing the function I(t), it is possible to say how quickly and which new articles reach their peak citation levels. For instance, plotting I(t) highlights peaks that correspond to highly influential periods of papers. Moreover, tracking the function R(t) lets monitor nodes that have lost their active influence. Studying the timestamps in which nodes transition into or out of R, gives the evaluation of how quickly various scientific articles lose relevance or regain popularity later. Finally, function S(t) tracks nodes that are not yet actively cited but have the potential to become influential in the future.

One of the questions in Project Phase B will be a practical examination of whether this method is suitable for accurately analyzing the research results. Unlike networks such as disease or infection spread, where changes occur rapidly, citation networks are assumed to be characterized by much slower dynamics. As a result, SIR model, which is well-suited for capturing fast changes, may not be appropriate for studying citation networks.

## **Approach to result analysis**

Result analysis is an important phase of the study. After obtaining the gravity models’ results for each node over different timestamps, the main task is to analyze it with a particular focus on how the influence of nodes changes over time. Result analysis can be in multiple forms and each gives a deeper understanding of the dynamic changes.

The first and the simplest way is **visualization of influence dynamics** **that involves creating graphs for each node**, where the x-axis represents timestamps, and the y-axis reflects the gravity model results. These graphs allow the most intuitive analysis of changes. Some nodes are represented by almost flat lines, often overlapping, while others by distinct curves that visually stand out. This approach provides an immediate way to identify interesting nodes. The nodes can be grouped into categories based on the graph behavior:

* Increasing influence - nodes whose influence is initially low but gradually grows over time;
* Decreasing influence - nodes whose influence is initially high but gradually reduces over time;
* Fluctuating influence - nodes whose impact has both increased and decreased over time;
* Stable influence - whose influence does not change over time.

Stable influence is less interesting for the research because it does not reflect dynamic or significant changes in the network.

The nodes with increasing influence are more interesting because a tendency indicates that these nodes are becoming more significant within the network. For example, if an article begins to gain influence, this could indicate that the researched topic is gaining popularity in the scientific community, thus opening up opportunities for further research toward the current direction.

The nodes with decreasing influence are also of interest. These nodes may point to articles that had a significant impact at one point but have become less relevant over time. This may happen for various reasons: the aging of information, the replacement of old theories with new ones, or changes in scientific priorities. Studying these nodes is important for understanding which ideas, theories, or studies have lost their relevance and why. This helps avoid unnecessary research in the current direction.

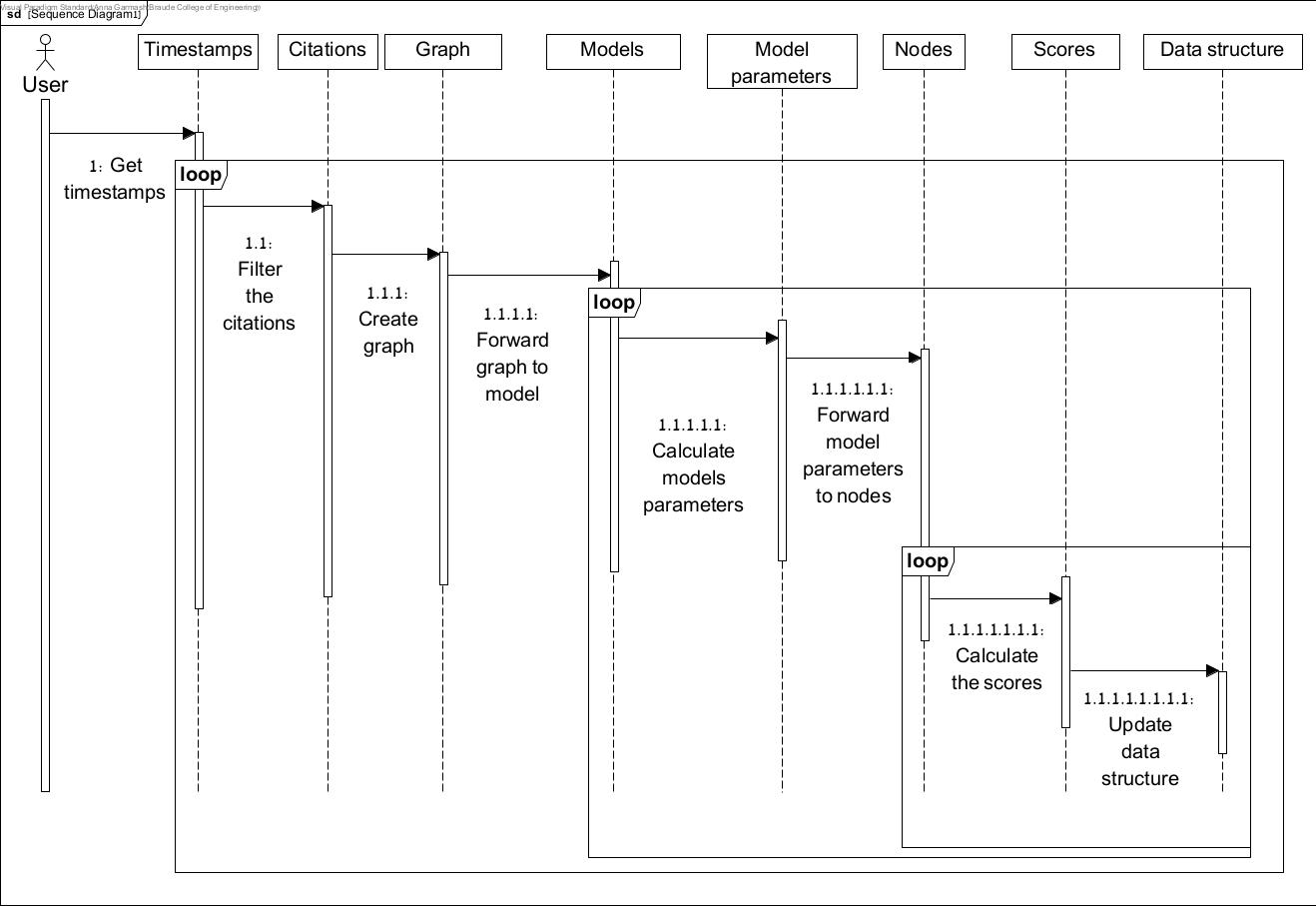
In this study, the examination of fluctuating influences also plays an important role because such graphs may point to areas of knowledge that are in a state of change or uncertainty. The reason for these fluctuations may include the influence of new competing topics, changing research interests, political or economic events, or even the discovery of errors or inaccuracies in research [[22](#bookmark=id.gb6asij2vj5j)]. In addition, articles whose influence has changed over time may indicate scientific topics that have not yet been fully accepted. Analyzing such nodes allows for trend prediction: whether the subject of the article might become more significant in the future or, conversely, lose its relevance.

Another approach is **comparing results across different timestamps**. Statistical methods such as Spearman’s rank correlation coefficient [[24](#bookmark=id.8p8b1k31gwyl)] or Kendall’s Tau coefficient [[25](#bookmark=id.uhwhf1bzuqmw)] can measure the degree of similarity or change in influence scores between one timestamp and another. If a correlation between timestamps is close to zero, it indicates changes in the network’s dynamics. Thus, it is possible to find out the period when the changes occurred and analyze which articles were added to the citation network. This reveals the articles that gain popularity, making them a key goal of research.

One more possible way of analysis is to **categorize nodes within each timestamp based on their level of influence** dividing them into "high influence," "medium influence," and "low influence" groups for example. The goal is to figure out whether nodes transition between categories. The number of categories can be adjusted as a parameter to find the most effective classification for analysis.

# **Related Diagram**

The following sequence diagram illustrates the flow in which the scores will be processed from the raw data to the designed data structure.



***Figure 22.*** *Sequence diagram.*

# **Testing plan**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Test** | **Input** | **Expected output** |
| Timestamps | Check if the dataset (citation network) has a publishing date for each document. | Dataset | Each document has dates. |
| Citation direction | Check if first column sites the second column in the dataset. | Dataset | Yes |
| Citation graph building | Check if the amount of the nodes in the graph equals to the number of the rows in the dataset in chosen timestamp. | Graph and dataset | Yes |
| Gravity Models | Check Gravity Models’ score correctness. | Gravity Model | Correct scores which capture local and global information. |
| Results visualization | Check if there is graph for each node and major lines are constant. Nodes with influence changes are visible. | 3D score matrix | Graph visualization for each node. |
| Results comparing across different timestamps | Check if the statistical methods results are in scope of predefined range. | 3D score matrix | Yes |
| Results comparing in nodes categorization | Check if the number of nodes across the groups are the same as the total number of rows in the dataset. | Groups and dataset | The number is the same. |

# **References**

[1] Newman, M. (2018). Networks. Oxford university press.

[2] Zhu, S., Zhan, J., & Li, X. (2023). Identifying influential nodes in complex networks using a gravity model based on the H-index method. Scientific Reports, 13(1), 16404.

[3] Bi, J., Jin, J., Qu, C., Zhan, X., Wang, G., & Yan, G. (2021). Temporal gravity model for important node identification in temporal networks. Chaos, Solitons & Fractals, 147, 110934.

[4] Duanbing Chen, Linyuan Lü, Ming-Sheng Shang, Yi-Cheng Zhang, Tao Zhou (2012). Identifying influential nodes in complex networks, Physica A: Statistical mechanics and its applications 391 (4) (2012) 1777–1787.

[5] Mark EJ Newman, A measure of betweenness centrality based on random walks, Social networks 27 (1) (2005) 39–54.

[6] Phillip Bonacich, Paulette Lloyd, Eigenvector-like measures of centrality for asymmetric relations, Social networks 23 (3) (2001) 191–201.

[7] L. Fei, Q. Zhang, Y. Deng, Identifying influential nodes in complex networks based on the inverse-square law, Physica A: Statistical Mechanics and its Applications 512 (2018) 1044–1059.

[8] Z. Li, T. Ren, X. Ma, S. Liu, Y. Zhang, T. Zhou, Identifying influential spreaders by gravity model, Scientific reports 9 (1) (2019) 1–7.

[9] F. Liu, Z. Wang, Y. Deng, Gmm: A generalized mechanics model for identifying the importance of nodes in complex networks, Knowledge-Based

Systems 193 (105464) (2020) 10.1016/j.knosys.2019.105464.

[10] H. Li, Q. Shang, Y. Deng, A generalized gravity model for influential spreaders identification in complex networks, Chaos, Solitons & Fractals 143 (2021),

[11] Na Zhao, Qian Liu, Hao Wang, Shenglong Yang, Pingzhou Li, Jian Wang, Estimating the relative importance of nodes in complex networks based on network embedding and gravity model, Journal of King Saud University - Computer and Information Sciences, Volume 35, Issue 9,2023,101758,ISSN 1319-1578,

[12] Cleanzr. (n.d.). GitHub - cleanzr/cora: Cora data set for Entity Resolution. GitHub. https://github.com/cleanzr/cora

[13] Citation Network Dataset. (2021, March 9). Kaggle. https://www.kaggle.com/datasets/mathurinache/citation-network-dataset/data

[14] Dirk Brockmann, Dirk Helbing, The hidden geometry of complex, network-driven contagion phenomena, science 342 (6164) (2013) 1337–1342.

[15] Zhao, J., Wen, T., Jahanshahi, H., & Cheong, K. H. (2022). The random walk-based gravity model to identify influential nodes in complex networks. Information Sciences, 609, 1706-1720.

[16] Qiuyan Shang, Yong Deng, Kang Hao Cheong, Identifying influential nodes in complex networks: Effective distance gravity model,Information Sciences,Volume 577,2021, Pages 162-179, ISSN 0020-0255,

[17] Hethcote, H. W. The mathematics of infectious diseases. SIAM Rev. 42, 599–653 (2000).

[18] Nisbet, R., Elder, J., & Miner, G. D. (2009). Handbook of statistical analysis and data mining applications. Academic press.

[19] Fan, W., He, Y., Han, X. *et al.* A new model to identify node importance in complex networks based on DEMATEL method. *Sci Rep* 11, 22829 (2021). https://doi.org/10.1038/s41598-021-02306-y

[20] Introduction to Social Media Investigation Jennifer Golbeck, 2015, 978-0-12-801656-5

[21] *Data Viz | Jacob Jameson*. (n.d.).

https://jacobjameson.com/ppha-30560/Final%20Project/

[22] Florides, G. A., & Christodoulides, P. (2023). Factors affecting the independence and reliability of Science and how these are perceived. SN Social Sciences, 3(2), 43.

[23] T-Test, Chi-Square, ANOVA, Regression, Correlation. . . (n.d.).

https://datatab.net/tutorial/pearson-correlation

[24] T-Test, Chi-Square, ANOVA, Regression, Correlation. . . (n.d.-c).

https://datatab.net/tutorial/spearman-correlation

[25] T-Test, Chi-Square, ANOVA, Regression, Correlation. . . (n.d.-d).

https://datatab.net/tutorial/kendalls-tau

[26] Kuckartz, U., Rädiker, S., Ebert, T., & Schehl, J. (2013). Statistik: eine verständliche Einführung. Springer-Verlag.

[GPT4]

[gpt1]<https://chatgpt.com/c/6778fbfb-bce4-800d-a00e-608488c0fc02>

[gpt2]<https://chatgpt.com/share/67794f35-8290-8013-8265-f839199fc450>